ABSTRACT OF THESIS

Name of Candidate Paul Andrew BIRO

Address 12A, Norfolk Road, LONDON, NW8 6HE

Degree Ph.D. *Date* June 1976

Title of Thesis Sequence Analysis of Mouse Satellite DNA.

The work described in this thesis concerns sequencing studies on mouse satellite DNA.

The satellite DNA was prepared from total mouse DNA by silver/caesium sulphate buoyant density gradient centrifugation. Radioactively labelled cRNA transcripts were prepared from both the native, double-stranded DNA and also from the isolated, heavy and light DNA strands, which had been separated on alkaline caesium chloride buoyant density gradients. The cRNA was subjected to fingerprint analysis after RNAase A and RNAase T_1 digestion.

Subsequently, EcoRII restriction fragments of the satellite DNA were hybridized to the isolated, light DNA strand and extended by DNA polymerase I, using ^{32}P-dnTPs to label the cDNA, which was then analysed by polyacrylamide gel electrophoresis, using the 'plus-and-minus' method of Sanger and Coulson, (J. Molec. Biol., 1975, 94, 441).

From these experiments, the sequence of the satellite was found to be based on a series of diverged, tandemly repeated, A-rich tracts, of which there were four major ones:- A_5UG, A_4CUG, A_4UG and A_3UG, which are all related and could have been derived from the same basic sequence. Sequence periodicities were observed within the satellite.

The evolution of the satellite DNA is discussed with respect to current theories on the evolution of repeated sequences and two possible models discussed in detail.

Use other side if necessary.

Sequence Analysis of Mouse Satellite DNA

by

Paul Andrew Biro

A Thesis Submitted to the

University of Edinburgh

for the Degree of

Doctor of Philosophy.

1976

ACKNOWLEDGEMENTS

This thesis describes work carried out by myself in the M.R.C. Mammalian Genome Unit at the Zoology Department, Edinburgh University, between October 1972 and October, 1975.

The thesis is my own work and the results presented in it are my own, except where otherwise stated. The first part of the work, concerning the cRNA sequencing, has been published in:- Biro, P.A., Carr-Brown, A.C., Southern, E.M. and Walker, P.M.B. (1975). J. Molec. Biol. 94, 71-86.

I wish to thank the M.R.C. for a Research Studentship.

I wish to thank my supervisor, Dr.E.M.Southern, for his invaluable assistance and also Professor P.M.B.Walker for much helpful advice, discussion and stimulation.

I would also like to thank Dr.K.Murray and the Staff of the Molecular Biology Department for advice and for gifts of various enzymes.

In addition, I wish to thank all at the M.R.C. Mammalian Genome Unit for their help in various ways, especially:- Alan, Barbara, Bob, Duncan, Fiona, Howard, Keith, Ken, John, Mary, May, Nick, Peter, Phil., Phoebe, Roger, Ron, Mrs.N.Buchanan and Mrs. M.Philp.

And also, Andrew, Colin and Pamela, John, John, Sandy, Eleanor, Hil., Nao., Graham, Gillian, Ann and Betty, among many others.

CONTENTS

CONTENTS (Cont.)

ABBREVIATIONS

DNA	Deoxyribonucleic acid
cDNA	complementary deoxyribonucleic acid
RNA	Ribonucleic acid
cRNA	complementary ribonucleic acid
PEI	Polyethylene imine
TEC	Triethylamine carbonate

Nucleotides:-

^{32}P-labelled triphosphates are labelled in the α position

A	Adenosine
C	Cytosine
G	Guanosine
T	Thymidine
U	Uridine

All sequences are written in the 5'→3' direction.

Restriction enzymes:-

Eco.RI, Eco.RII.	Restriction endonucleases from *E.Coli.* with the RI and RII plasmids respectively
Hind.II, Hind.III	Restriction endonucleases from *Haemophilus influenzae*, serotype D.
Hpa.I, Hpa.II.	Restriction endonucleases from *Haemophilus parainfluenzae*.
Hae.III.	Restriction endonuclease from *Haemophilus aegyptius*.

SUMMARY

The work described in this thesis concerns sequencing studies on mouse satellite DNA.

The satellite DNA was prepared from total mouse DNA by silver/caesium sulphate buoyant density gradient centrifugation. Radioactively labelled cRNA transcripts were prepared from both the native, double-stranded DNA and also from the isolated, heavy and light DNA strands, which had been separated on alkaline caesium chloride buoyant density gradients. The cRNA was subjected to fingerprint analysis after RNAase A and RNAase T_1 digestion.

Subsequently, Eco.RII restriction fragments of the satellite DNA were hybridized to the isolated, light DNA strand and extended by DNA polymerase I, using ^{32}P-dnTPs to label the cDNA, which was then analysed by polyacrylamide gel electrophoresis, using the 'plus-and-minus' method of Sanger and Coulson, (J. Molec. Biol., 1975, 94, 441).

From these experiments, the sequence of the satellite was found to be based on a series of diverged, tandemly repeated, A-rich tracts, of which there were four major ones:- A_5UG, A_4CUG, A_4UG and A_3UG, which are all related and could have been derived from the same basic sequence. Sequence periodicities were observed within the satellite.

The evolution of the satellite DNA is discussed with respect to current theories on the evolution of repeated sequences and two possible models discussed in detail.

CHAPTER ONE

Introduction

The DNA of an organism contains all the genetic information required to specify the protein and RNA species which it makes. However, eukaryotes contain more DNA than would seem to be required solely to code for these proteins and RNA species. For example, the DNA content of a typical mammal is approximately $5x10^9$ base pairs (b.p.), (Sparrow et al. 1972), which is sufficient to code for roughly $5x10^6$ proteins, each approximately 300 amino acids long. Although estimates of the number of structural genes which code for proteins vary considerably (from $0.5x10^5$ to $5x10^5$ genes), not more than 10% of that DNA would be required even for $5x10^5$ proteins, (Kimura and Ohta, 1971; Rosbash et al. 1975).

Furthermore, the actual amount of DNA present in different species varies considerably, even between closely related species, which presumably have a similar number of gene products. This is the so-called 'C-value' paradox, (Ohno, 1972; Edström and Lambert, 1975). For example, the toad has 11 pg. of DNA per diploid cell, whereas the newt has 86 pg. of DNA per cell, (Bachman et al. 1972).

Little is known about the function, or even the true extent, of this apparently 'extra' DNA. Some of it may be involved in control elements or may have a structural role in maintaining chromosome morphology, (Ohno, 1972; Lewin,1974). However, some of this DNA is known to be present in the form of relatively homogeneous, tandemly repeated, short sequences. Such sequences provide an easily accessible fraction of DNA as their base composition may be sufficiently different from that of the bulk of the DNA for them to be prepared by methods such as buoyant density gradient centrifugation.

Such sequences have been termed 'satellite' sequences, (Walker, 1971a) and are almost certainly not required for coding functions. An examination of them would provide information on the function and organization of DNA in chromosomes and on the evolutionary developement of the genetic material.

Physical properties of satellite DNA.

In 1957 Meselson et al. discovered that native DNA could be fractionated on neutral caesium chloride (CsCl) density gradients, as the buoyant density of a given DNA under such conditions is directly proportional to its G+C content.

Using this technique, Kit (1961) examined the DNA from a variety of animals and found a light component in mouse DNA which had a buoyant density significantly less than that of the bulk of the DNA, (1.691 g/ml. as opposed to 1.701 g/ml., McConaughy and McCarthy, 1970). Similar satellite components, both heavy and light, have since been detected in a large variety of both animals and plants, (Arrighi et al. 1970; Coudray et al.1970; Walker et al. 1969; Ingle et al. 1973).

Mouse and other satellites can also be isolated in prep--arative amounts by buoyant density gradient centrifugation using a fixed angle rotor which Flamm et al. (1966) showed had a 10x higher capacity for DNA and gave greater separation of DNA than a swinging bucket rotor, first suggested by Fisher et al. (1964). Although CsCl centrifugation has been widely used to detect and prepare such satellites, heavy metal ion / caesium sulphate gradients offer several advantages over the use of CsCl alone, (Corneo et al. 1968a). They have a higher capacity for DNA and

give greater separation of DNA components. Although the theory is not fully understood, at the appropriate alkaline pH (usually just over pH 9) heavy metal ions bind preferentially to either AT-rich (Ag^{+}) or to GC-rich (Hg^{++}) DNA and so considerably alter the buoyant densities of such sequences over that of the bulk of the DNA. Consequently, greater separation is obtained than by CsCl alone, (Corneo et al.1970a).

Corneo et al. (1968a) used Ag^{+}/Cs_2SO_4 gradients to prepare mouse satellite and this is now the usual method, (see, for example, Cech et al. 1973). Skinner and Beattie (1973) used Cs_2SO_4 gradients containing both Hg^{++} and Ag^{+} ions to separate two crustacean satellites which form a single band in CsCl gradients, 'isopycnic twin' satellites. Buoyant density centrifugation has also been used to isolate non-satellite sequences such as the repeated ribosomal genes in amphibians (Birnstiel et al. 1968).

A variety of antibiotics are available which also can either bind to AT-rich regions (Netropsin) or to GC-rich regions (Actino--mycin) and have, for example, been used by Peacock et al. (1973) to isolate Drosophila satellites and by Brown et al. (1971) to prepare amphibian ribosomal genes. More recently, Paprooka and Müller (1974) have shown that commercially produced dyes such as malachite green or phenylated neutral red can bind to specific DNA sequences and consequently be used to fractionate DNA. Other techniques which do not involve the use of gradients have also been used, for example, Cheng and Sueoka (1961) used methylated albumin kieselguhr columns to isolate crab satellite.

The properties of isolated satellites have been extensively studied and a considerable amount of information is available on

their structure (see reviews by Walker, 1971a; Rae, 1972; Jones, 1973; McGregor, et al. 1973 and also Hennig et al. 1970).

The amount of satellite present varies greatly between different organisms, even between closely related members of the same family. In the spider crab, Libinia, the poly d(AT) satellite comprises 6% of the genome, in Cancer irratus it comprises 10% of the genome and in Cancer borealis 30%, whereas in other crabs it may be entirely absent, (Sueoka, 1961; Skinner et al. 1970). Similarily, in the kangaroo rat, Dipodomys spectabilis, it is hardly detectable, whereas in Dipodomys ordii it constitutes more than 50% of the DNA, (Mazrimas and Hatch, 1972). A species may also contain more than one satellite; for example, the Guinea pig has three; the ox at least two, (Corneo et al. 1970a) and man has at least three, (Corneo et al. 1970b; Jones et al. 1974).

The mouse satellite has been studied by several groups, (Flamm et al. 1966b; Corneo et al. 1968a; and the information is summarized in Walker, 1971a). It has a buoyant density of 1.691 in neutral CsCl and bands as a light satellite as mouse main band DNA has a buoyant density of about 1.701. It comprises 10% of the mouse genome and contains 50% G+T and 35% G+C. However, this figure of 35% is a little higher than that predicted from its buoyant density in CsCl, (31%) and lower than that predicted from its thermal denaturation profile, (41%), (Corneo et al. 1968a). Furthermore, Flamm et al. (1966b and 1967) showed that reassociated satellite had a higher density than the native satellite and that when reassociated satellite is again melted the buoyant density decreased but not to the same extent as the native satellite. These anomalies between the expected and the observed properties of the satellite,

which have also been found in the Guinea pig satellite I, (Corneo et al. 1968a) are probably due to the arrangement of G/C and A/T base pairs in the sequence which will alter the properties of the satellite over a random polymer of the same base composition (Wells et al. 1970). The anomalous reassociation behaviour of the satellite may be caused by mismatching of diverged sequences in the reassociated duplex, (Flamm et al. 1966b). These results show how the sequence of a stretch of DNA may influence and alter its properties.

Flamm et al. (1967) showed that the two strands of the satellite DNA could be separated by alkaline CsCl density gradient centrifugation as their G+T contents are sufficiently different. At an alkaline pH of 12.5 only G and T bases are titrated and hence only they contribute to the increase in buoyant density caused by alkali, (Vinograd et al. 1963; Flamm, 1972). The heavy strand of the satellite has 58% G+T and a density of 1.752 g/cm^3 at pH 12.5, whereas the light strand has 42% G+T and a density of only 1.725 g/cm^3 at pH 12.5, (Flamm et al. 1967). This property, which is shared by many satellites, for example the Guinea pig satellite I, (Corneo et al. 1968a; Flamm et al. 1969b), human satellites I and II, (Corneo et al. 1968b), Drosophila satellites, (Gall et al. 1973) and ox satellites I and II, (Corneo et al. 1970a), but not by all, for example satellites from various Apodemus species, (Walker et al. 1969), is very useful for sequencing as it enables each DNA strand to be analyzed separately and without contamination by the other. It has been utilized by several workers, for example, Southern, (1970), Gall and Atherton, (1974) and is also used in the analysis of the mouse satellite, described below.

The suggestion that satellites might be associated with the centromeres of chromosomes was first made by Schildkraut and Maio, (1968), who showed that the DNA from purified mouse nucleoli was enriched in satellite. Yasmineh and Yunis, (1969 and 1970) examined DNA from differently sedimenting fractions of chromatin and found that the centric heterochromatic fraction contained nearly 70% of the satellite DNA and the remaining 30% was located in an intermediate fraction containing both euchromatin and heterochromatin. Similar enrichment has been found in other systems, for example, in Guinea pig and in ox (Yunis and Yasmineh, 1970 and 1971, respectively) and in kangaroo rat, (Mazrimas and Hatch, 1970).

Confirmation of the localization of mouse satellite sequences was made by the technique of in situ cytological hybridization, (Gall and Pardue, 1969; John et al. 1969; Buongiorno-Nardelli and Amaldi, 1970) who hybridized radioactively labelled nucleic acids to standard cytological chromosome smears which had been treated with mild denaturing agents to separate the complementary DNA strands, (Pardue and Gall, 1975).

Pardue and Gall, (1970), used ^{3}H-labelled mouse satellite DNA and Jones, (1970) used ^{3}H-labelled mouse satellite cRNA to show satellite sequences were located at or near the centromere of all mouse chromosomes, except possibly the Y chromosome. This has been confirmed in other systems, for example the fly, Rhinchosciara, (Eckhart and Gall, 1971), for Drosophila, (Rae, 1970; Peacock, 1973), for kangaroo rat, (Prescott et al. 1973) and for the ox satellites which hybridize to all except the sex chromosomes which also have been shown by C-banding to lack constitutive heterochromatin, (Kurnit et al. 1974), and is probably a general

phenomenon for satellites, (Rae, 1972). Although Hennig et al. (1970), also found satellite sequences in the euchromatin of Drosophila hydei but concluded these were probably sites of interstitial heterochromatin. Such centromeric regions of chromo--somes are composed of highly repetitive DNA sequences, but they may not all appear as satellites after buoyant density centrifug--ation if their base composition is not sufficiently different from that of the bulk of the DNA, (Rae, 1970; Yunis and Yasmineh, 1972).

The distribution of satellites among chromosomes may vary, for example, the human satellites, I, II, and III, are not uniformly distributed among human chromosomes. Satellite II is predominantly on nos. 1,9 and 16, (Jones and Corneo, 1972), but satellite I is mainly on the 1 and 3 chromosomes and satellite III is mostly on the no. 9, although all three satellites are present on almost all the chromosomes, at least to a limited extent, (Jones et al. 1974).

Heterochromatin is known to be late replicating in the 'S' phase of the cell cycle, (Lima de Faria, 1969), and it has been shown by several workers using pulse labelling of synchronized cells that satellites also replicate late in the 'S' phase. Tobia et al. (1970) labelled synchronously growing mouse L cells with both ^{14}C and ^{32}P, followed by buoyant density centrifugation, to show that DNA with a high G+C content replicated early in 'S' phase and DNA with a low G+C content replicated late in 'S', ie. satellite-like DNA. Flamm et al. (1971) used a mouse lymphoma cell line and found that the satellite itself was late replicating. A similar result was obtained by Bostock and Prescott, (1971) who

used ^{3}H-thymidine as label, although the satellite may not be the last component to be replicated in the 'S' phase. Similarily, Bostock et al. (1972) also showed that the kangaroo rat satellite was late replicating. These results support the view that the satellite is an integral component of constitutive heterochromatin, and the function of heterochromatin may be the function of satellite.

Reassociation

When a solution of DNA is heated above its melting temperature (Tm), it will denature and, if allowed to cool slowly, the denatured duplex will reassociate to form a double helix again, (Marmur et al. 1963). In 1961 Marmur and Doty, (1961), found that whole, simple DNAs, for example from phages, reassociated more rapidly than DNA from more complex organisms such as bacteria. However, this is only true for unique, (non-repeated), DNA and Britten and Kohne, (1966) sheared DNA from various organisms to a uniform size class, as the rate of reassociation is proportional to the size of the DNA, and showed that the rate of reassociation of any particular DNA is inversley proportional to its complexity. They defined the complexity of DNA as the amount of diverse DNA sequence in a given DNA preparation. If repeated sequences are absent it will be equal to the genome size, if repeated sequences are present it will be a measure of the length of the individual repeating unit, (see also Bolton et al. 1965).

For a number of years reassociation was an important technique for studying the genome of eukaryotic organisms and a picture was built up of the DNA from such organisms being composed of three principal components. Firstly, there was a slow renaturing

component, (unique DNA), an intermediate component and a fast renaturing component, (repetitive DNA), (Britten and Smith, 1970; Walker, 1971a). The unique DNA would comprise individual genes present in only one copy, the intermediate DNA multiple copy genes and control sequences, and the fast renaturing component would comprise repetitive DNAs such as satellite. Renaturation characteristics led Britten and Davidson to produce their hypothesis on gene regulation, (1969).

The factors which influence the rate at which denatured DNA will reassociate have been extensively studied, for example, by Britten and Kohne, (1966), Britten, (1969) and by Wetmur and Davidson, (1968) who showed it was a second order reaction with a temperature optimum 25°C below the Tm. Wetmur and Davidson proposed a model involving a two-step process - an initial pairing of complementary, denatured DNA strands, (nucleation), followed by a fast 'zippering' together of the strands to form the renatured helix. The initial nucleation is sequence dependent and is rate determining and hence the reaction is second order.

When mouse satellite DNA is reassociated the reaction proceeds very quickly indeed; it is in the fast renaturing component of the DNA. Waring and Britten, (1966) estimated the $Cot_{\frac{1}{2}}$ value to be 6.6×10^{-4}. (Reassociation rates are measured in terms of $Cot_{\frac{1}{2}}$ values. Cot is the product of the initial DNA concentration and the reassociation time to any given percentage of reassociation. Hence $Cot_{\frac{1}{2}}$ is the value of Cot at which the DNA is 50% reassociated. The lower the $Cot_{\frac{1}{2}}$ value, the more rapid the reassociation). Britten and Kohne, (1968) estimated the value to be 10^{-3}. For

comparison, T_4 phage has a $Cot_{\frac{1}{2}}$ of 1: hence mouse satellite reassociates 10^3 times as fast as T_4 phage DNA. On this basis the complexity of mouse satellite DNA is 300 base pairs, (b.p.), and it is present in at least 10^6 copies per genome, (Waring and Britten, 1966).

The value of 300-400 b.p. for the complexity does not agree with the estimate of Southern, (1970), of 12 b.p. which was based on pyrimidine tract analysis of the satellite and assumed that all tracts which were of the same general composition were related and were equivalent. This kind of discrepancy was first noted for the Guinea pig satellite I. Sequence studies had shown it was based on a hexanucleotide repeat, (Southern, 1970) but reassociation data, (Corneo et al. 1970a) suggested a complexity of 10^5 b.p. and those of Sutton and McCallum, (1971) a value of 100 b.p. after correction for mismatching which is still a greater figure than the 6 b.p. found by sequencing.

Three hypotheses were put forward to explain these discrepancies. The first suggests that it is due to the effects of mismatching. When a short, tandemly repeated, diverged sequence is allowed to reassociate there will be some mismatching of the base pairs in the reformed heteroduplex. This mismatching is known to affect its stability, (Waring and Britten,1966; Laird et al. 1969; Gelderman et al. 1971). Britten and Kohne, (1968) quote a value of 1°C reduction of Tm for every 1% of bases mismatched and Laird et al. (1969), estimate the same reduction of Tm for every 1.4% mismatching. Reassociated mouse satellite is thermally less stable than native satellite by 5°C, (Flamm et al .1967), hence considerable mismatching does occur when satellite is reassociated. Southern, (1970)

suggested that the mismatching could affect the rate of reassociation as well as the stability of the resultant reassociated duplex.

Southern, (1971), considered that the initial nucleated duplex, formed as the first stage in the reassociation reaction, might be destabilized by mismatched bases and this could have a pronounced effect on the subsequent observed rate of reassociation. Hence the sequence complexity, as measured from the kinetic data, which is in effect the kinetic complexity, could be an over-estimate of the true, sequence complexity, if indeed there is a large effect. for example, Britten and Smith, (1970) estimated the kinetic complexity of calf satellite I to be 17,000 b.p. which Southern corrects to a proposed sequence complexity of only 600 b.p.

Since that time other analyses of satellites have supported this view. Allan, (1974) measured the rate of reassociation of Apodemus agrarius satellite and showed that mismatch can reduce the reassociation rate until the satellite appears to have the same complexity as T_4 phage DNA, ie. has a $Cot_{\frac{1}{2}}$ of 1, although further experiments, (quoted in Southern, 1974) have indicated it has a sequence complexity of only 300 b.p.

Sutton and McCallum, (1971), randomly sheared mouse satellite DNA and fractionated it into different classes according to thermal stability by melting the reassociated duplex off a hydroxylapatite column at progressively increasing temperatures. The more stable sequences would then elute at a higher temperature. When each of four of such classes was melted and allowed to renature, there was a clear correlation between reduction in the resultant $Cot_{\frac{1}{2}}$ value and reduction in the extent of mismatching. The more closely matched sequences reassociated more rapidly. They estimate the kinetic complexity of the satellite at 300 b.p., in agreement with

other published values, (Waring and Britten, 1966), but reduced it to 120 b.p. when mismatching was taken into account. In a subsequent paper, Sutton and McCallum, (1972) further suggested that the complexity may be as low as 10-20 b.p. after they cross-reassociated M.mus with M.caroli satellites as in this case there seemed to be a much shorter reassociation register, (see section on the evolution of the satellite).

However, such serious effects of mismatching have not always been found by other workers. Lee and Wetmur, (1973) and Hutton and Wetmur, (1973) used chemically modified bases to investigate their effect on reassociation rates. Lee and Wetmur found that a 13°C reduction of Tm was equivalent to 11.5% mismatching and caused a twofold reduction in the rate of reassociation. Hutton and Wetmur used deaminated and glyoxylated Lambda phage DNA and E.coli DNA and found that for deaminated DNA, 33% mismatching gave a 23°C reduction of Tm and a twofold reduction of reassociation rate. Glyoxylated DNA had a 17°C lowering of Tm when 16% mismatched and also gave a twofold reduction of reassociation rate. They conclude that for small differences in Tm, less than 8°C, such as found by Sutton and McCallum, the reassociation rate would not be altered by more than 20% and that the two to three-fold differences found in the mouse satellite were caused by differences in the ancestral sequences between the four fractions.

McCarthy and Farquar, (1972) deaminated the C residues of B.subtilis DNA by alkali treatment to form deoxyU residues and, when they measured the reassociation rate with increasing mismatching, found that a fourfold reduction in rate could be caused by as little as 9% mismatching. They suggest that, for repeated sequence

DNA, values of the complexity obtained from kinetic analysis will be overestimated.

Bonner et al. (1973), cross-reassociated related bacterial DNAs and also reassociated deaminated E.coli and phage DNAs and also found a large effect of mismatching on the reassociation rate, although not as great as Sutton and McCallum or McCarthy and Farquar found. Bonner et al. estimate that 10% mismatching causes a 10°C reduction of Tm and a twofold decrease in the reassociation rate. They point out that their estimate is only one-third of Sutton and McCallum's and propose that mouse satellite may have a faster and a slower reassociating component, (Cf. Marx and Hearst, below).

The second hypothesis to account for the anomalous renaturation kinetics of mouse satellite DNA was suggested by Hutton and Wetmur, (1973b). They measured the reassociation rates of several classes of short lengths of satellite, (from 580±100 b.p. down to 36±9 b.p.) which had been produced by the action of S1 nuclease, which is specific for single strands of DNA, on denatured mouse satellite and found that the rate increased with decreasing size of DNA fragment. They estimated the maximum complexity of the satellite was 36±9 b.p. which was based on the smallest size class of DNA used in their experiments and suggested that it could be as low as Southern's estimate of 12 b.p. (1970). They proposed a model for the satellite in which satellite sequences 36 b.p., or less, in length were separated from one another by variable, non-satellite regions. They concluded that reassociation kinetics cannot be used to estimate complexities of repeated sequences as the reassociation rate will depend only on the size

of the DNA used.

Chilton, (1973) has proposed a theoretical model to account for this effect. Although the rate of reassociation of unique DNA is directly proportional to the molecular weight, this is not the case for repeated sequence DNA. The rate is independent of the repeat size (ie. of the complexity) and is proportional only to the inverse of the square root of the molecular weight, which is a kinetic parameter for diffusion controlled reactions, (Schmitz and Schurr, 1972). Hence increasing the size of the DNA leads to a reduction in the reassociation rate and values for the complexity of DNA derived from such data only reflect the size of the DNA used in the experiments.

The third hypothesis suggests that mouse satellite may be composed of more than one component, each of which would have different reassociation rates. This was first suggested by Hutton and Wetmur, (1973a) to account for the appreciable differences in reassociation rates observed by Sutton and McCallum, (1971), and has also been advanced by Bonner et al. (1973). Recently, Marx and Hearst, (1975) used mouse satellite which had been sheared to 410 b.p. and found it reassociated as a two component system, by performing the reaction in 0.36M. Na^+ and also in 0.075M. Na^+. They estimate the complexity of the fast component to be 130 ± 25 b.p., which agrees with that of Sutton and McCallum, and of the slow component as $1,250\pm50$ b.p. and the mole fractions as 0.69 and 0.31, respectively.

Since the estimates of Sutton and McCallum and of Marx and Hearst agree with that derived from restriction analysis (Southern, 1975) of 120 b.p. it would appear to be possible to derive a value

for the complexity of repeated sequence DNA by using kinetic data, although the result obtained would have to be corrected for the effects of mismatching and, even then, may not represent the shortest sequence periodicity, but only the register in which the reassociation reaction is occurring under those conditions. Other techniques for obtaining estimates of complexities include sequencing and restriction analysis.

Restriction analysis

The availability of sequence specific endonucleases, restriction enzymes, (Smith and Wilcox, 1970; Kelly and Smith, 1970), has provided a powerful means of investigating the structure and long range sequence organization of DNA, (Southern and Roizes, 1973; Hörz et al. 1974; Nathans and Smith, 1975). In contrast, sequencing examines the short range structure of DNA.

Mouse DNA was among the first to be investigated by this technique. Botchan et al. (1973) restricted total mouse DNA and analyzed the fragments so produced by agarose gel electrophoresis. They showed that the satellite ran as a slow moving band, behind the bulk of the DNA and was, therefore, largely refractory to the enzymes used which were:- Eco.RI, Hpa.I and Hpa.II endonucleases, which indicates that the basic satellite sequence does not contain the ~~restriction~~ recognition sequences of any of these enzymes and that the overall sequence is non-random, as otherwise several restriction sites for these enzymes would have been found in the satellite DNA.

Similarily, Hörz et al. (1974) restricted the purified satellites of the Guinea pig and found that although most of the DNA was unaffected by the enzyme used, Hind.II + III, satellites I and III, but not II, produced a restriction pattern of a few

faint bands.

The structure of calf DNA has been similarily investigated by Philippsen, (1974), and by Mowbray and Landy, (1974), who showed that the satellite produced bands which could be observed over the bulk of the DNA after restriction. Botchan, (1974) restricted the purified major satellite, (satellite I), which comprises 6-7% of the total genome, (Kurnit et al. 1973), with Eco.RI and showed that the major band obtained was 1,400 b.p. in length and proposed that the satellite consists of repeating units of that length which had originally been produced by a multiplication process, such as by unequal crossing-over or by a rolling circle mechanism. Philippsen et al.(1975), have now shown that the 1,400 b.p. unit is, in fact, repeated throughout the satellite.

Subsequently, Mowbray et al. (1975), performed a more detailed analysis using Hind.II + III and Hae. as well as Eco.RI and extended Botchan's model by postulating a possible decanucleotide sequence as the basic repeating unit which would have given rise to the restriction sites by divergence and which has subsequently been multiplied to produce different hierarchies of repeats. Similar investigations are currently being carried out on Drosophila satellites, (Manteuil et al. 1975).

Southern, (1975) restricted mouse satellite DNA with Eco.RII and found that it produced a series of bands of which the major component comprised 70% of the total satellite DNA and was approximately 240 b.p. in length and the other components were multiples of this length and were present in progressively decreasing amounts. Consequently, he proposed a repeating unit of 240 b.p. (the monomer) for the satellite with multiples being

produced by loss of the intermediate restriction sites, owing to divergence. In addition to the major pattern, a second, fainter series of bands was observed half-way between the major bands, ie. at 120, 360, 600 b.p. and suggesting that the 240 b.p. monomer is itself composed of two 120 b.p. units which lack the intermediate restriction site. The minor pattern of fragments may have been produced by unequal crossing-over within the 240 b.p. register.

Southern eluted the three major bands, the monomer, dimer and the trimer, denatured them and allowed them to self-reassociate, whereupon each size class produced a series of longer, staggered reassociated duplexes as well as ordinary monomers, dimers and trimers. The reassociation register appeared to be half the length of the monomer, ie. 120 b.p., and thus showed that the two 120 b.p. units which make up the monomer are of similar sequence.

Although the satellite is largely undigested by another restriction enzyme, Hae.III, a small proportion of the DNA is susceptible to it and produces a series of bands similar to that produced by the Eco.RII enzyme. The Hae.III restriction site may have been introduced into the sequence by divergence and spread by unequal crossing-over to adjacent monomers and so have formed an identifyable sub-family of the satellite.

Hörz et al. (1974), restricted the satellite with Hind.II + III and, although the DNA was also largely unaffected, a minor restriction pattern, similar to that produced by the Hae.III enzyme, was found and the intervals between the bands were estimated at 225 b.p.. Hörz proposed that the Hind. restriction sites were introduced into thesequence by a similar mechanism to the Eco.RII sites, into approximately 4% of the satellite and have been lost

by divergence. Roughly 70% of the Hind. sites must have mutated in order to give rise to the observed restriction pattern, which corresponds to 20% divergence between the monomers. However, Southern estimates that only 3% divergence is necessary to account for the observed Eco.RII restriction pattern. It is, presumably, possible that some satellite components, containing the Hind., and also the Hae.III sites, have a higher divergence rate than the bulk of the satellite which also contains the Eco.RII sites, or are older and have therefore accumulated more mutations. However, a simpler explanation would be that these sites have been spread by unequal crossing-over, as Southern suggests, (1975).

Hörz has pointed out that the satellite DNA is preferentially methylated; it contains 4% methylcytosine, (Salomon et al. 1969), and suggests that methylation, rather than divergence or unequal crossing-over, could also account for the restriction patterns. Mouse satellite contains 17.5% C, (Walker, 1971a), and, therefore, over 20% of all C bases will be methylated. However, this would not be sufficient to account for the loss of 70% of all the Hind. sites, unless they are preferentially methylated, although it would eliminate about 60% of the Eco.RII sites which is considerably more than the 14% which Southern estimates have been lost, unless the Eco.RII sites are preferentially under-methylated. Most methylation of eukaryotic DNA is in CpG couplets and Harbers et al. (1975), have shown that 40% of all methylated C bases occur as purine-C-purine, (A/G-C-A/G) in mouse satellite DNA, which sequences are not often found in the restriction sites. Hence methylation is non-random and , although it could account for the loss of some of the sites, it is unlikely to account for all

the observed restriction patterns.

Function of satellite DNA

Despite some earlier indications of Harel et al. (1968), there is now no evidence that satellite DNA is transcribed in vivo. Flamm et al. (1969), hybridized RNA which had been extracted from mouse liver, kidney, spleen and myeloma tissue, to ^{32}P-labelled, separated heavy and light satellite DNA strands, in the ratio 10^4 parts RNA : 1 part DNA and would consequently have detected 1 in 10^4 hybrid molecules on hydroxylapatite. However, no hybrid molecules were found. They also hybridized excess ^{32}P-labelled RNA to DNA and would have detected 1 in $4x10^3$ hybrid molecules, but, again, failed to detect any hybridization.

Furthermore, sequence analysis has shown that the simple, short repeats found in satellites could only code for a very limited range of proteins, (Southern, 1970). Finally, their location in the genetically inert centromeric heterochromatin, (Yasmineh and Yunis, 1970), also suggests that satellites are not involved in protein synthesis and are not transcribed in vivo.

A number of possible functions have been suggested for satellites, (Walker, 1971a and b; Bostock, 1972; Britten, 1972; Flamm, 1972; Yunis and Yasmineh, 1972; and Davidson and Britten, 1973).

Walker, (1968, 1969 and 1971b), has proposed a series of 'housekeeping functions' for satellites. For example, he argues that their location at the centromere of chromosomes could promote Robertsonian fusion between non-homologous chromosmes which have the same satellite sequences, as there seems to be a measure of 'Affinity' between certain non-homologous chromosomes of the same

origin, (Wallace, 1958). Such fusion processes could aid in speciation which was first suggested as a possible function for centric heterochromatin by Darlington, (1937). However, as Robertsonian fusion is a rare process, Bostock, (1972) argues that this is an unlikely function for satellite.

A second possible housekeeping function for satellite is in promoting chromosome recognition of homologous chromosomes for meiotic and mitotic pairing, (Walker, 1971b; Walker et al. 1969). However, the apparent random distribution of satellites between chromosomes, whereby the same satellite may be present on all or only a few chromosomes, makes this an unlikely possibility. Although Kurnit et al. (1974), find that the bovine sex chromosomes lack all the four bovine satellites and suggest that there may be a link between meiotic pairing of chromosomes, (although not sex chromosome function), and satellite distribution.

Walker, (1968), proposed a further 'housekeeping' function in which the satellite regulated the folding pattern of the chromosome during mitotic condensation. However, the location of the satellite at the centromere, rather than in the euchromatic regions which are actually involved in the folding process, also provides evidence against this suggestion, (Bostock, 1972).

Walker, (1971b), has discussed the possibility that satellite is concerned with the phenomenon of 'centromeric strength', whereby a 'strong centromere', which would contain more centric heterochromatin can pull a 'weaker centromere' to its pole in double first anaphase bridges, (Lindsley and Novitski, 1958). Sederoff et al. (1975) have suggested that differential distribution of satellite sequences may be responsible for centromeric strength in

Drosophila melanogaster.

Several authors have proposed that repeated sequences which are located in the euchromatic regions may be sites for genetic control, for example by acting as sites for the initiation of transcription or of translation or by acting as regulatory sequences, (Britten and Davidson, 1969; Georgiev, 1969; Davidson and Britten, 1971). It has also been suggested that satellite sequences of some organisms may be located in such euchromatic regions, (Hennig et al. 1970; Arrighi et al. 1970). Although there is no in situ cytological evidence, Flamm et al. (1969), estimated, on the basis of hybridization experiments, that about 1% of the mouse main band DNA could interact with satellite DNA. As a control they found that heterologous DNA, such as rat or Guinea pig DNA, contained no mouse satellite sequences at all. Consequently, in those cases, satellites which were located in euchromatic regions could act as recognition sites for recombination or as sites for the initiation of transcription, as suggested by Walker, (1969).

Guille and Quetier, (1973), have put forward a detailed model along such lines. They suggest repeated genes could activate the genes which lie adjacent to them. Coarse control would be exerted by the number of repeated sequences adjacent to the genes and fine control by metabolic modification of the sequences, such as by methylation and point out that mouse satellite is preferentially methylated, (Salomon et al. 1969), and hence could act in this capacity.

However, it seems hard to accept the view that the only satellite sequences to have any function would be those which are located away from the bulk of the satellite, which is located in the heterochromatin, the function of which would still remain unsolved.

Sutton, (1972), has put forward the view that DNA can cause chromatin 'crystalization' as the conformation of any given stretch of chromatin will depend on the DNA sequence involved and 'like' conformations will then crystalize to form a more condensed structure. The strength of the packing forces will depend on the unusualness of the repeated chromatin conformation and the frequency and similarity of the repeated sequences involved. Hence satellite would promote heterochromatinization and unique sequence DNA would favour a loose chromatin structure. In this way satellite would be involved in both chromosome structure and in coarse genetic control.

Kram et al. (1972), had suggested that repeated sequences in the centric heterochromatin of Drosophila melanogaster may be organized in segments, about 3×10^6 daltons long, and interspersed with non-repeated sequence DNA which, if it were active in transcription, could be under the control of the repeated sequence DNA. However, Peacock et al. (1973) found no evidence for such interspersed sequences after isolating satellite DNA up to 20×10^6 daltons in weight.

Britten and Davidson, (1971) suggested that satellites may constitute pools of non-coding DNA which would be free of selection pressures and therefore able to evolve rapidly to give rise to new functional DNAs, such as protein coding sequences. There is evidence that heterochromatin itself may evolve more rapidly than euchromatin. Dev et al. (1975) compared the Q- and C-band staining patterns of Mus musculus musculus with those of Mus musculus molossinus, (two different strains of Mus musculus), and showed that the Q-banding patterns, which are found in the euchromatic

regions of the chromosomes, were identical in the two strains, whereas the C-banding patterns which are located in the heterochromatic regions, were different. There is also evidence that non-coding DNA evolves more rapidly than coding DNA. McCarthy and Farquhar, (1972), cite the sequence differences between rDNA genes in Xenopus mulleri and Xenopus laevis as examples. (The genes themselves have identical sequences in the two species, but the non-transcribed spacers differ by approximately 14% of their sites, (Brown et al. 1972; see also Discussion section). Although satellite DNA evolves at the same rate as unique DNA, (Southern 1974a), Rosbash et al. (1975) have shown by cross-hybridizing mouse mRNA to rat single-copy DNA, and vice-versa, that non-coding sequences evolve more rapidly than coding sequences. Approximately 1% of repeated sequences will be lost by divergence every 10^6 years, (Rice, 1972; Southern, 1974a), and hence there is loss of satellite which provides additional support for the Britten and Davidson hypothesis.

Mazrimas and Hatch, (1972), have suggested that satellites confer greater genetic flexibility on the organism, as they find that the DNA of more highly specialized kangaroo rat species contains less satellite than the more primitive species. However, Skinner et al. (1973), point out that the distribution of poly d(AT) satellites in the Crustacea appears random throughout the Crustacean families and cannot be correlated with specialization or with taxonomic position, (Skinner, 1970). Instead, Skinner suggests that satellites may be involved in chromosome recognition.

Any suggested function would have to take into account all

the known properties of the satellite. To date there is no single proposed function which does this. However, a mechanism must exist for the formation of repeated sequences such as satellites and, presumably, they must confer some selective advantage to the organism for them to be formed and maintained over a period of evolutionary time. Although, as pointed out above, there will be a slow loss of satellite from rodents, Gall and Atherton, (1974), have shown that Drosophila virilis satellites are very homogeneous and hence may be highly conserved, more so than rodent satellites. There may, therefore, be mechanisms such as unequal crossing-over, which maintain sequence homogeneity, at least in some systems. Indeed, Sutton, (quoted in Walker, 1971b) has argued that satellite sequences may be maintained as long as they remain genetically inert and so, for example, do not initiate transcription, but may be lost when divergence has eroded this property.

Evolution of satellites

The presence of a large number of related, tandemly repeated sequences suggests a common origin for all of them.

Britten and Kohne, (1969), proposed that satellites arose by a sudden multiplication of an original ancestral sequence - saltatory replication. Evidence for sequence periodicities within the satellite, (Southern, 1975; Mowbray et al. 1975; and to be discussed in this thesis), would argue against a single multi--plication event, but in favour of a more gradual process, possibly a series of such events, over a longer period of time.

However, satellites must have arisen comparatively recently, as even closley related species may contain different ones. Walker et al. (1969) showed that purified, isolated single strands

of mouse satellite DNA failed to hybridize to rat DNA which, therefore, does not contain mouse satellite sequences, although it does contain repeated sequences, (Rice, 1972), which do not form a satellite on buoyant density centrifugation, (Kit, 1961; McConaughy and McCarthy, 1970). Walker et al. subsequently examined the satellites of five different rodent species by analytical ultra-centrifugation and showed that they had very different buoyant densities, suggesting they had different sequences. Hennig and Walker, (1970), went on to examine the satellites from other, related rodent species, including Microtus and Arvicola from the Cricetidae family and Rattus, Mus and Apodemus from the Muridae, and, again, showed that they exhibited considerable differences in buoyant density and rates of reassociation.

Some closely related species are known to contain similar satellites. Hennig et al. (1970), cross-hybridized the satellites from two Drosophila species, (neohydei and pseudoneohydei), which can inter-breed and concluded that they were related but distinct. Skinner and co-workers have examined many Crustacean satellites by both cross-hybridization and by ultra-centrifugation, (Skinner, 1967; Beattie and Skinner, 1972) and concluded that each satellite was distinct. Graham and Skinner, (1973), demonstrated that the homology between satellites decreases with the evolutionary divergence of the species which suggests that the satellites may have had a common origin and have since diverged and also that the satellites are relatively young, possibly only as old as the species itself.

Jones et al. (1973), found that human satellite III can cross-hybridize with chimpanzee and with orang-outang DNA and Prosser et al. (1973) found a chimpanzee satellite which can

cross-hybridize with human DNA which indicates that, in each case, the satellites had common ancestors. Sutton and McCallum, (1972), examined the satellites of four closely related mouse species, Mus caroli, M. cervicolor, M. famulus and M. mus, by cross-hybridization and showed they were all closely related and had probably diverged from different elements of a highly diverged common ancestral sequence. Southern, (1975), demonstrated the existence of a faint Eco.RII restriction band from M. caroli DNA which was equivalent to the monomer band found in M. mus DNA after treatment with the same enzyme, and suggested that although most of the multiplication of M. mus satellite occurred after the two species had diverged, some sequences in M. caroli derived from an ancestor to the M. mus satellite. Indeed, Rice, (1972) found by cross-hybridization that these two Mus species do contain some common repeated sequences.

Many organisms contain more than one satellite, which may or may not be related. The Guinea pig contains at least three, (Corneo et al. 1968a, 1970a), one of which, the satellite I, is a heavy satellite and is not related to the other two which are light satellites and which may be related, as suggested by Southern, (1972), on the basis of pyrimidine tract analysis. Skinner et al. (1973), have sequenced the major satellite, (I), of the hermit crab, Pagurus, and have shown that it is not related to the minor satellite, (II). Such non-relationships would be expected if the satellites in question arose from completely different DNA sequences.

Conversely, in some species, notably Drosophila, satellites are found which do share a common origin. Blumenfeld, (1973), has cross-hybridized isolated single strands of the three Drosophila

virilis satellites and showed that they were closely related and that they too shared a common origin. This has been supported by the sequencing studies of Gall and co-workers, (Gall et al. 1973; Gall and Atherton, 1974). Similarily, Peacock et al. (1973) and Endow et al. (1975), have sequenced the Drosophila melanogaster satellites, (see section on 'sequencing', below), and showed relationships between them.

The work to be described below is concerned with the sequence and the organization of the mouse satellite DNA. The sequencing data will be discussed with special reference to the evolution of the satellite which will be shown to be consistent with the model that satellites evolved from a simple ancestral sequence by divergence and multiplication. The nature of the original sequence will be discussed and possible stages in the multiplication of the satellite suggested. Finally, the mechanism of satellite developement will be discussed in relation to current theories on the evolution of repeated sequences.

CHAPTER TWO

Sequence Analysis of Mouse Satellite cRNA

Introduction.

Sequence analysis forms a direct approach to the examination of individual satellites and provides information on their short-range structure, their relatedeness and also on their evolution.

Until recently, the only specific degredative method of direct examination of DNA has been the analysis of the pyrimidine tracts which are produced after diphenylamine degredation of DNA, (Burton and Petersen, 1960; Ling, 1972). This method has been used by Southern, (1970), to determine the principal sequence of the Guinea pig satellite I and, subsequently, to examine the satellites II and III, Southern, (1972; 1974a). It has also been used by Peacock et al. (1973) to compare the sequences of three of the five Drosophila melanogaster satellites. Salamon et al. (1969) and, more recently, Harbers and Spencer, (1974), and Harbers et al. (1974) have applied it to mouse satellite and their results have confirmed and extended those of Southern, (quoted in Walker, 1971a, and published in Biro et al. 1975).

However, most satellite sequencing work to date has involved the use of a cRNA transcript. RNA sequencing methods, such as used in this analysis are now well established, (Sanger et al. 1965; Brownlee and Sanger, 1967, 1969; Sanger and Brownlee, 1967; Adams et al. 1969), and have been reviewed, (Brownlee, 1972; Murray, 1974).

Generally, the DNA is transcribed in vitro with E. coli RNA polymerase, (Burgess, 1969), and is radioactively labelled at one or more specific bases, usually with ^{32}P, by adding the appropriate $\alpha^{32}P$-labelled nucleoside triphosphate to the incubation medium.

After purification, the cRNA may be sequenced by the usual techniques of site-specific digestion, followed by fingerprint analysis.

Such methods have been used to sequence a number of different satellites. Fry et al. (1973) sequenced the HS-β satellite of the kangaroo rat, Dipodomys ordii, and found it was based on a diverged decanucleotide:- ...ACACAGCGGG..., although its extent of divergence was less than the Guinea pig or the mouse satellites. Gall and Atherton, (1974), have sequenced the three satellites of Drosophila virilis and showed they were all heptanucleotides which were related to one another by simple base changes. Satellite I is related to satellite II by one C to T change:- ...ACAAACT... to ...ATAAACT..., and to satellite III by a different C to T change:- ...ACAAACT... to ...ACAAATT.... The three satellites are very homogeneous, being no more than 1% mismatched. Similarily, Endow et al. (1975) have studied three of the four major satellites of Drosophila melanogaster and have shown that they, too, are highly conserved, are related to one another and can be described by the general formula:- $(AA-)_m(A-)_n$. Satellite II is a decanucleotide:- ...AATAACATAG... and is highly conserved: 95% of the DNA is present as one sequence. Satellite II may be a twelve nucleotide repeat containing 2(AAT):3(AT) and satellite IV may be based on a pentanucleotide:- ...AAGAG.... However, they are not as highly conserved as the Drosophila virilis satellites.

Skinner et al. (1973) separated the strands of the major satellite, (I), of the hermit crab, Pagurus pollicaris, and sequenced each transcript separately. The satellite was shown to be based on a tetrameric repeating unit:- $...\frac{CCTA}{GGAT}...$ which makes up 90% of the sequence, and a further 5% is made up of a minor,

unrelated sequence which forms a separate satellite, (II). Although there do not seem to be any obvious sequence relationships between those satellites, from different orders, whose sequences are known, Skinner et al. propose that the Guinea pig satellite I sequence:- CCCTAA, (Southern, 1970) is related to the CCTA of the hermit crab and that they may have had a common evolutionary origin. Recently, Salser, (unpublished results) has shown that the basic sequence of the HS-α satellite of the kangaroo rat is the same as that of the Guinea pig satellite I, also suggesting a common evolutionary ancestor.

In the analysis of the mouse satellite, presented below, the DNA strands were separated on alkaline CsCl density gradients, transcribed with the E. coli RNA polymerase and the products analyzed by conventional two-dimensional fingerprinting techniques.

Materials and Methods

DNA Preparation

Mus musculus DNA was prepared by a modification of the method of Walker and McClaren, (1965). Mice were starved overnight to deplete the livers of glycogen, killed and the livers and, in some cases, the testes and kidneys as well, were removed, homogen--ized and subjected to standard procedures of phenol extraction, followed by chloroform/octanol extractions, and ethanol spooling. After redissolution, the DNA was treated with ribonuclease and, subsequently, pronase and then further extracted with phenol and chloroform/octanol. Final purification consisted of glycogen pelleting, followed by one or two isopropanol spoolings. Care was taken to avoid the use of chloride-containing buffers as subsequent

steps involved the use of silver-containing gradients which would have formed AgCl precipitates with any chloride ions present. Approximately 1mg. of purified DNA was obtained from each mouse. The purity of the DNA was confirmed spectrophotometrically. I am also grateful to Miss M. White for gifts of purified mouse DNA.

Satellite was prepared from total mouse DNA by silver/caesium sulphate buoyant density gradient centrifugation, (Corneo et al. 1968a), using an Ag^+/phosphate ratio of 0.27. The solution, containing up to 4mg. DNA/18ml. gradient, was centrifuged for a minimum of 48hr. at 32,000r.p.m. in an M.S.E. 8x40 titanium fixed angle-head rotor at 20°C. Gradients were fractionated from the top by aspiration using a plastic float, satellite fractions pooled and subjected to a second cycle of centrifugation if required to achieve complete purification. Silver was removed by dialysis against 3M. NaCl solution, followed by dialysis against dilute Tris-HCl/EDTA (10mM pH 7.4) (0.1mM) buffer and the DNA concentrated by pelleting and redissolved in water and stored at 4°C. Satellite prepared by this method was found to be essentially free of main-band contamination by analysis in an M.S.E. analytical ultracentrifuge.

Separate, heavy and light, strands of the satellite DNA were prepared on an alkaline CsCl gradient by centrifugation in 10ml. tubes placed in adaptors in an M.S.E. 8x40 titanium rotor for 20hr. at 40,000r.p.m. at 20°C. (Flamm et al. 1967). Gradients were fractionated by piercing the tube and collecting fractions from the bottom. About 25µg. of DNA could be handled on each gradient, and the caesium removed by dialysis against dilute Tris-HCl/EDTA buffer and the DNA concentrated by evaporation and stored frozen at -20°C.

cRNA Preparation.

RNA polymerase holoenzyme was made by the method of Burgess, (1969), using the DE52 column and glycerol gradient procedures which retain the sigma factor. The purified enzyme was found to be free of contaminating RNAase, DNAase and polynucleotide phosphorylase activities by appropriate assays.

Transcription of the satellite DNA was carried out in a maximum volume of 20µl per 1µg DNA. The reaction mixture comprised:- 40mM Tris-HCl, (pH 7.9), 100mM $MgCl_2$, 0.1mM EDTA, 0.1mM dithio--threitol, 150mM KCl, 0.05% bovine serum albumin and 0.15mM UTP, GTP, CTP, and ATP, one, or more, of which would carry a suitable radioactive label, and a typical incubation would contain 5µg DNA and 10µl of enzyme in 80µl of reaction mix. Transcription would be carried out for 30min at 37°C on a polythene sheet covered by a plastic cap and sealed with vacuum grease or in an 'Eppendorf' plastic reaction tube. Later it was found that a higher yield of RNA could be obtained by using a higher initial nucleotide concentration, (0.3mM), of each and adding a second 10µl of enzyme after 30min. and allowing the reaction to proceed for a further 25min. Radioactive nucleotides were purchased from the Radiochemical Centre, Amersham or from New England Nuclear Corporation. 0.4µCi of each of ^{3}H-ATP, ^{3}H-CTP, ^{3}H-GTP, and ^{3}H-UTP were added to the incubation mixture for the synthesis of ^{3}H-labelled cRNA. The specific activity of the ^{14}C-GTP was 500mCi/mmol and of the ^{32}P-ATP and ^{32}P-UTP was approximately 2Ci/mmol and of the ^{32}P-CTP and ^{32}P-GTP was 100Ci/mmol.

0.2µg of RNAase-free DNAase I, (Worthington), was then added per 1ug of DNA present and incubation continued for a further 25min. The mixture was diluted to 2ml with 0.1M NaCl, 0.2% sodium

laury sulphate, extracted three times with buffer-saturated phenol, (0.3M sodium acetate, 2mM EDTA, pH 7.0), and the cRNA precipitated with at least three volumes of absolute ethanol at -20°C for a minimum of 4hr. After pelleting, the cRNA was dissolved in a minimal volume of water, (50-100µl), and desalted on a Sephadex G75 or G100 column which was eluted with water. Fractions containing the cRNA were pooled, dried under vacuum and stored on polythene sheets at -20°C.

Characterization of the cRNA.

(I) Sizing

A sample of the cRNA, labelled with ^{32}P-ATP was run on a 5% polyacrylamide gel, (Loening, 1967), together with some total myeloma RNA, (a gift of Dr.M.L.Sartirana) to provide standards of known mobility. The gel was scanned on a Joyce-Loebl ultraviolet scanner, sliced and counted by the Čerenkov method.

(II) Hybridization

(a) DNA excess.

25µg of satellite DNA was run on a standard alkaline CsCl gradient and 0.25ml fractions collected. 0.3ml of 6XSSC, (1XSSC is 0.15M sodium chloride, 0.015M sodium citrate), was added to each and 0.2ml samples diluted into 2.0ml buffer, (1.5M NaCl, 0.5M Tris-HCl, pH 7.9), which were poured over 'Millipore' nitrocellulose filters, (HAWP 0.45µm), which had been thoroughly washed in 2XSSC and set up in 'Millipore' filter holders. The filters were then further washed with 2XSSC, dried at room temperature for 4hr and subsequently baked under vacuum at 80°C for 2hr.

0.5ml of ^{3}H-labelled cRNA, (120×10^{3}cpm at a specific activity of 80×10^{3}cpm/µg RNA), was boiled for 5min to denature the RNA,

rapidly cooled by dilution into 8ml of 2XSSC and 0.5ml of the solution added to each of the filters in glass phials. Hybridization was carried out at 60°C for 1¼hr. The filters were then washed for 1hr in two changes of 2XSSC at room temperature, dried at 37°C and counted in toluene-based scintillation fluid.

After removal from the scintillant, the filters were washed in toluene and treated with 15ml of heat-treated RNAase A, (10µg/ml) for 1hr at 37°C, washed in 2XSSC, dried and recounted, (Gillespie and Spiegelman, 1965).

(b) RNA excess.

2µl of each of the satellite fractions in 6XSSC were each further diluted into 0.5ml samples of 6XSSC, each of which contained 1µg of carrier Micrococcus lysodeikticus DNA, and bound to nitrocellulose filters in the usual way.

1ml of the ^{3}H-labelled cRNA, (240×10^{3}cpm), was boiled, diluted into 5.5ml of 2XSSC and 0.4ml of solution added to each filter. Hybridization was carried out for 3.5hr at 60°C, after which the filters were washed, treated with RNA~~ased~~, dried and counted.

Fingerprinting.

Approximately 0.5×10^{5}cpm of each ^{32}P-labelled cRNA, (or ^{14}C-labelled cRNA), was digested with either 0.5µg RNAase T_1, (Sankyo), or 0.5µg RNAase A, (Worthington), in 10µl of incubation mixture, (10mM Tris-HCl, 1mM EDTA, pH 7.6), containing 10µg E. coli tRNA to give the correct enzyme:RNA ratio, (1:20). The reaction was carried out on polythene sheets, covered with a plastic cap and sealed with vacuum grease. Optimum incubation times varied from 30 to 60 min. After rapid drying under vacuum, the digested cRNA was taken up in 2µl of buffer for electrophoresis

on cellulose-acetate strips, (Oxoid, Ltd.), 95cm x 2cm, (Sanger and Brownlee, 1967). For RNAase T_1 digestion products the buffer used was 7M urea, 5% acetic acid, adjusted to pH 4.3 with pyridine and electrophoresis was carried out for 3.5hr at 50V/cm. For RNAase A digestion products the same buffer was used, but the pH was left unadjusted at around pH 3.5 and electrophoresis was for 2hr at 50V/cm. The oligonucleotides were then allowed to transfer to PEI-impregnated cellulose thin layers, (Southern and Mitchell, 1971) by capillary action, (Southern, 1974b).

PEI-cellulose thin layers were prepared by adding cellulose powder, (MN300,Camlab), to a 1% PEI solution adjusted to pH 6.5 with formic acid, homogenizing the solution, degassing it and spreading the slurry thinly over plastic sheeting using a metal applicator. 30g of cellulose in 200ml of solution were sufficient for 5,000cm^2 which made three layers, each approximately 40cm x 40cm on Kodak X-ray film backing. The layers were stored at 4'C for at least two weeks to allow them to dry and harden. Immediately before use they were wetted, washed in 2M formic acid adjusted to pH 2.3 with pyridine for 10min, then in water for a further 10min, and allowed to air dry until only slightly damp.

After transferring the oligonucleotides, the layers were soaked in water to wash off the urea, blotted dry and developed in pyridine-formate solvent, using a filter-paper wick clipped to the top of the layer to absorb the solvent. RNAase T_1 digestion products were developed until the blue marker dye had moved about 30cm, (two-thirds of the way to the top of the layer), which took about 16hr. The solvent used was 1.8M formic acid adjusted to pH 3.8 with pyridine. RNAase A digestion products were chromatographed in 2.0M formic acid adjusted to pH 3.5, (Ford

and Southern, 1973).

After drying, oligonucleotides were located by radioautography, cut out, counted in toluene-based scintillator, washed in toluene and dried. ^{14}C-GTP labelled oligonucleotides were also counted in a gas-flow counter. The PEI-cellulose, containing the oligonucleo--tides, was wetted and scraped into the plastic tops of small syringe needles which had been lined with filter-paper to cover their holes. Elution was carried out with 3x10µl portions of 2.0M triethyl-ammonium-carbonate which were squeezed through the needle onto polythene sheets and allowed to evaporate under partial vacuum.

Each sample was divided into two, one half was subjected to nearest-neighbour analysis by hydrolysis with 5µl of 0.5M NaOH in sealed glass capillaries for at least 24hr at 37°C. The digestion products were separated by electrophoresis on either Whatman no.1 or no.52 paper for 2hr, at 60V/cm, using a 5% acetic acid, 0.5% pyridine buffer, (pH 3.5).

The other half was subjected to secondary digestion: RNAase T_1 products with RNAase A, and RNAase A products with RNAase T_1. In each case, the oligonucleotides were taken up in 5µl of 10mM Tris-HCl, 1mM EDTA, (pH 7.6), containing 5µg carrier *E. coli* tRNA and 0.25µg of enzyme and incubated for 30min at 37°C. 2µl of 0.5M HCl was then added to each sample and incubation continued for 60min to destroy cyclic nucleotides.

Separation and identification of the products of secondary digestion was carried out by electrophoresis on DE81 paper in 5% acetic acid, 0.5% pyridine, (pH 3.5),(Adams et al. 1969), and also on DE81 paper in 7% formic acid to separate the large A-rich oligonucleotides.

Results

Pyrimidine Tract Analysis of Mouse Satellite DNA.

Preliminary sequence analysis of the satellite had been undertaken in this laboratory by A.Carr-Brown and E.M.Southern who had examined the pyrimidine tracts of ^{32}P-labelled satellite DNA on DEAE-cellulose columns, (Burton and Petersen, 1960; Petersen and Reeves, 1966) and had subsequently sequenced the major tracts by labelling non-radioactive tracts with ^{32}P at their 5' ends using polynucleotide kinase, (Szekely and Sanger, 1969; Murray, 1973), followed by partial exonucleolytic digestion with venom phosphodiesterase. The products of digestion were separated and identified by electrophoresis on AE-paper, (Southern, 1970). The results have been published, (Biro et al. 1975; quoted in Walker, 1971a) and suggested that the mouse satellite, like the Guinea pig satellite I, is a diverged sequence, based on a fairly simple repeating unit. Preliminary analysis of these results suggested that the sequence might be approximately 12 b.p. long, (Southern, 1970; Walker, 1971a).

The principal tracts from the heavy strand were T_4C, T_5C and similar sequences which could be related to them by one or two base changes, and which can be described by the general formula:- T_nC. They are in good agreement with those published by Harbers and Spencer, (1974), and by Harbers et al. (1974), although the latter authors suggest that the evolution of the satellite may not be as straightforward as Southern, (1970; Walker, 1971a) suggests. The results obtained of the pyrimidine tract analysis are shown in Table 1.

TABLE I

Pyrimidine tracts of mouse satellite DNA strands

data of A.Carr-Brown and E.M.Southern

Isoplith	R.L.	Compositional isomer	R.L.	Principal Sequences present
Heavy strand				
1	12	C	40	
		T	15	
2	40	C_2	120	
		C,T	80	TC, CT
		T_2	250	
3	160	C_3	3,000	
		C_2,T	450	CTC, TCC, CCT
		C,T_2	450	CTT, TTC, TCT
		T_3	700	
4	44	C_4	-	
		C_3,T	500	
		C_2,T_2	75	TCCT, TTCC, TCTC
		C,T_3	155	TTTC, TTCT
		T_4	900	
5	82	C_3,T_2	1,200	TTTTC, TTTCT
		C_2,T_3	300	TTTCC, TTCTC
		C,T_4	130	
6	33	C_4,T_2	850	
		C_3,T_3	130	TCCTTC, TTCCTC

TABLE I (cont.)

Isoplith	R.L.	Compositional isomer	R.L.	Principal Sequences present
6 (cont)		C_2,T_4	60	TTTCTC, TTTTCC
		C,T_5	120	TTTTTC, TTTTCT
		T_6	450	
7	140	C_2,T_5	-	TTTTCTC
		C,T_6	-	TTTTCTT, TTTTT(CT)
8	160	C_3,T_5	-	TTTTC(TC_2)
9-13	160			
Light strand				
1	3.5	C	14	
		T	17	
2	15	C_2	270	
		C,T	40	CT, TC
		T_2	120	
3	47	C_3	380	
		C_2,T	130	TCC
		C,T_2	100	CTT
		T_3	350	
4	125	C_3,T	850	
		C_2,T_2	600	
		C,T_3	330	CTTT

TABLE I (cont.)

Isoplith	R.L.
5	130
6	250
7	320
8	514

R.L. = Repeat Length

The average repeat lengths of each tract were calculated from the formula:-

$$\frac{100(n-1)}{X},$$

where 'n' is the number of bases in the tract, 'X' is the proportion of ^{32}P in the tract, measured by column chromatography, and is expressed as a percentage. The average repeat length is then the average distance, in base pairs, between tracts of the same length and/or base composition. The accuracy of measurements of the compositional isomers decreases for the rarer oligonucleotides.

Transcription of Native Satellite DNA.

The next stage of the analysis involved the preparation and characterization of a cRNA transcript suitable for subsequent sequence analysis.

A sample of purified ^{32}P-ATP labelled cRNA was sized by polyacrylamide gel electrophoresis, using total myeloma RNA as a marker, (Fig.1). The cRNA was found to be heterodisperse with an average length of around 120 nucleotides: very little was smaller than 4s. The size of the RNA would affect the results as, assuming that the start and the end-point of the transcription is random, small, non-representative oligonucleotides will be produced from sequences at the 3' and 5' ends of the RNA after site-specific cleavage. Such oligonucleotides will appear in the fingerprints as increased random background of spots, thus making the sequence appear to be more complex than it actually is. However, as the transcript is 120 bases long, this effect will not be too serious.

A sample of ^{3}H-labelled cRNA was hybridized to separated single strands of satellite DNA to determine the symmetry of transcription, (Gillespie and Spiegelman, 1965).

The hybridization was first performed under conditions of excess DNA, (Fig.2). In previous experiments, in which the hybridization had been carried out for 12hr, there appeared to have been some loss of DNA from the nitrocellulose filters, probably in the form of well-matched RNA/DNA duplexes, and this resulted in a trough in hybridization at the peak of the DNA. In order to prevent this, hybridization was only carried out for $1\frac{1}{4}$hr, rather than the 3hr which would be required for saturation, assuming that the satellite had a complexity similar to that of

Fig.1 Sizing of cRNA by gel electrophoresis

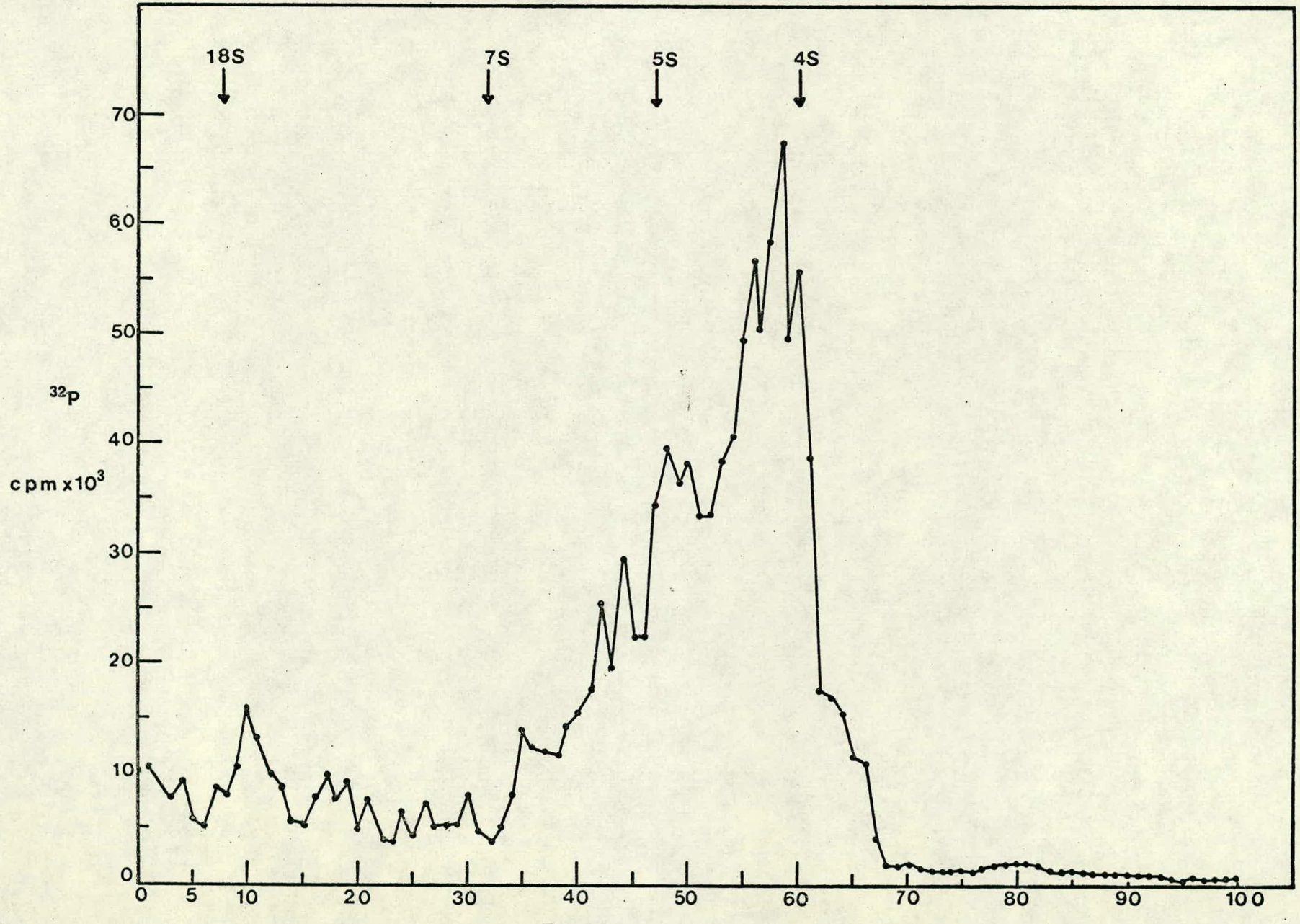
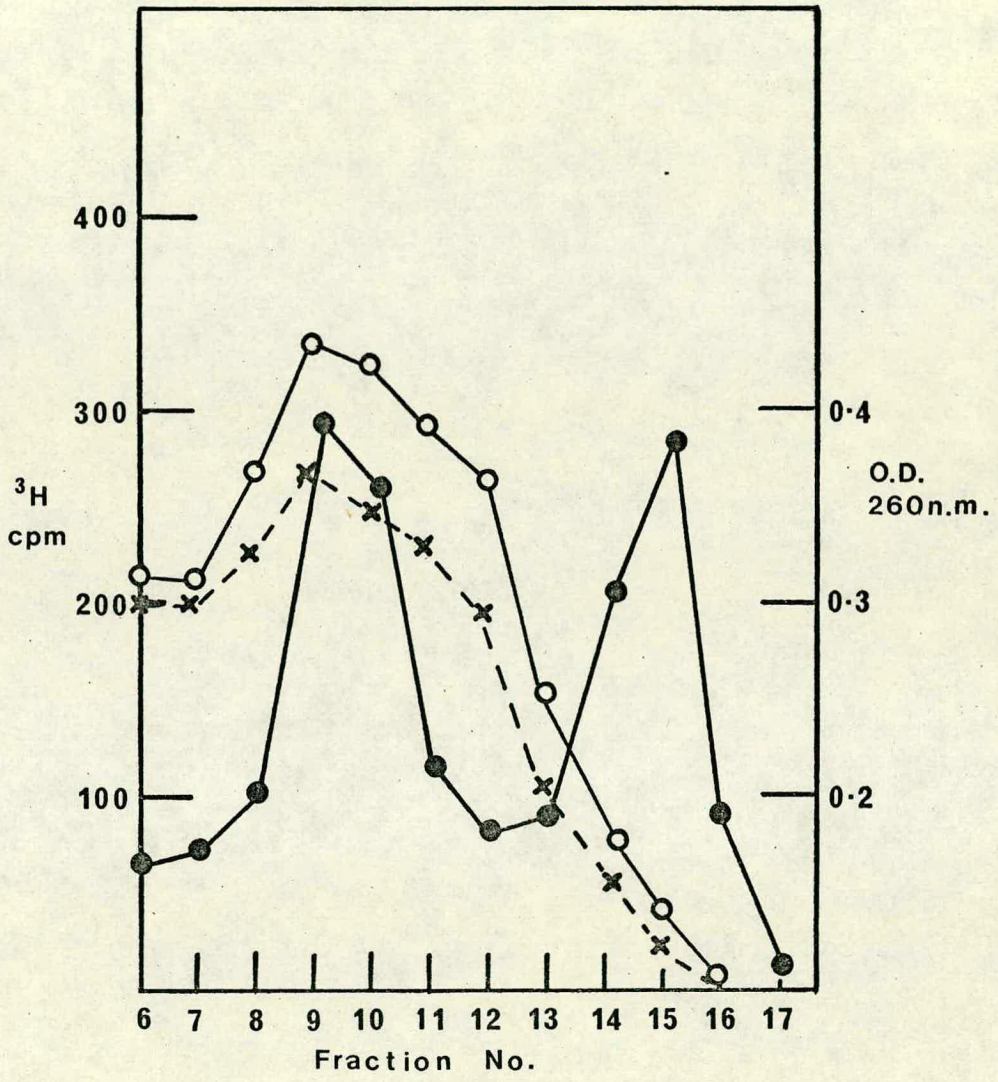


Fig.2 DNA Excess Hybridization



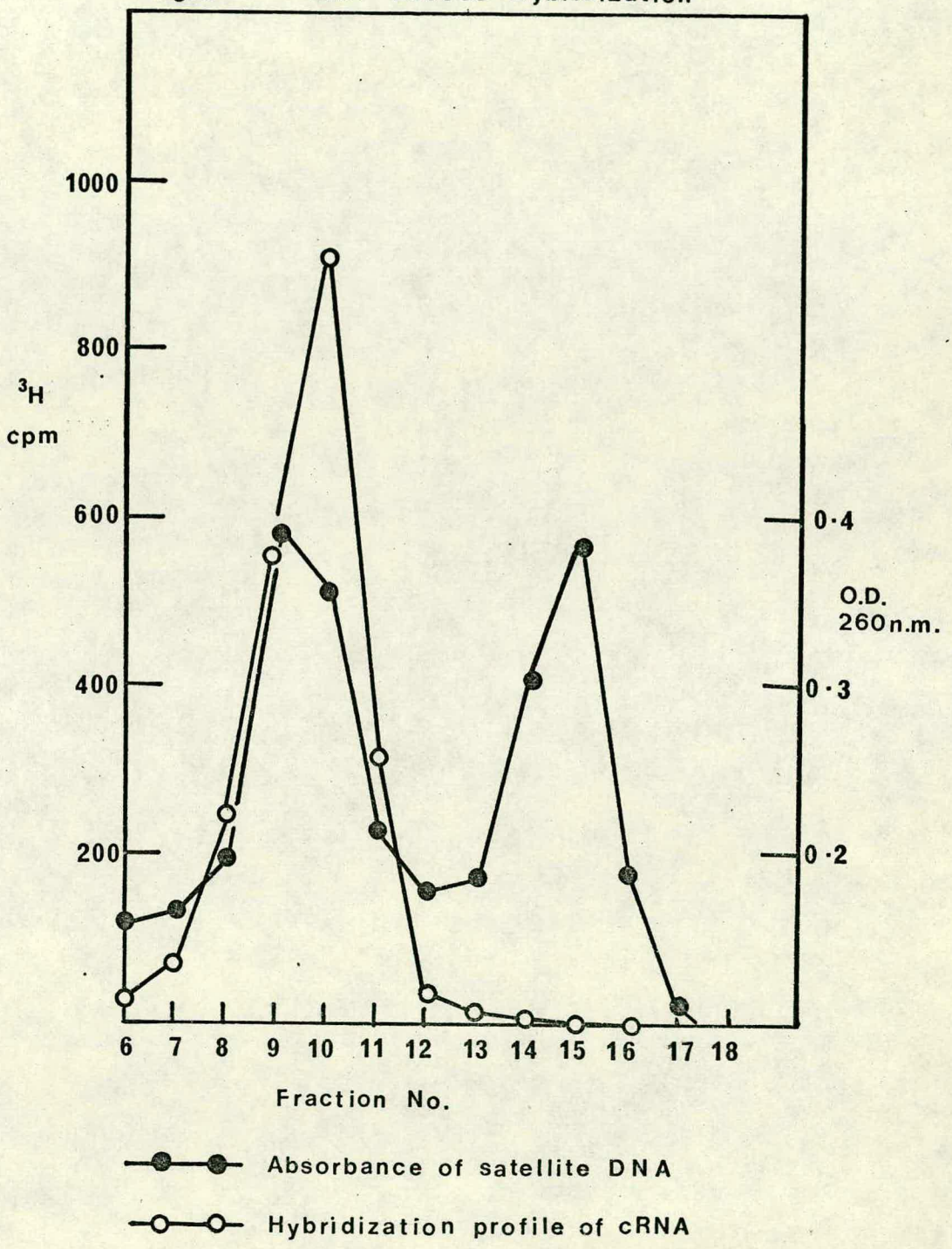
—●—● Absorbance of satellite DNA

Hybridization profile of cRNA:-

—○—○— Before RNAase treatment

-×-×- After RNAase treatment

Fig.3 RNA Excess Hybridization



Xenopus tRNA: about 120 b.p., (Birnstiel et al. 1972), which is now known to be the reassociation register of the satellite DNA, (Southern, 1975). The hybridization profile, shown in Fig. 2, shows that the cRNA is a copy of the heavy, pyrimidine-rich DNA strand; no hybridization at all took place on the light strand.

The experiment was repeated under conditions of 50-100 fold excess RNA, (Fig.3) and a similar result was obtained: the cRNA only hybridized to the heavy DNA strand. Hence the cRNA produced from native satellite DNA was exclusively a copy of the heavy strand and produced fingerprints which were identical to those obtained from cRNA transcribed from the isolated heavy strand.

RNA polymerase isolated from E. coli has been observed to transcribe other satellites asymmetrically, for example Drosophila melanogaster satellite IV, (Endow et al. 1975), but it is not a universal phenomenon as it transcribes both strands of Apodemus species' satellites, (Allan, 1974), although it does have a preference for one strand. Similarily, Brownlee et al. (1974) have observed that it copies only one strand of Xenopus 5s DNA.

Quantitative Analysis of the cRNA.

An estimate of the sequence complexity of an RNA species can be obtained by an examination of the fingerprints produced after site-specific cleavage with an enzyme such as RNAase T_1. The number of different digestion products, their relative positions, and their relative amounts, all supply information on the sequence complexity.

cRNA was synthesized from isolated heavy and light strands of the satellite DNA, using ^{14}C-GTP as radioactive label, and fingerprints made from the RNAase T_1 digestion products. RNAase T_1

cuts RNA on the 3' side of each guanine residue and so each oligonucleotide in a complete digest contains only one guanine which will be at its 3' terminal. Thus the proportion of ^{14}C radio--activity in each oligonucleotide is a measure of its mole fraction in the sequence. The same information cannot be so accurately obtained when α^{32}P-GTP is used to label the RNA as oligonucleotides which have an extra guanine residue adjacent to the one at the 3' end will contain two ^{32}P groups after RNAase T_1 treatment owing to phosphate transfer. Such oligonucleotides will appear to be present in twice their actual amounts and hence double their mole fraction. Conversely, the label will be lost altogether from the isolated guanine residue which will not appear in the fingerprint as a visible spot.

The fingerprints produced show a relatively simple pattern, (Plates I and II). The heavy strand transcript, ($cRNA_H$), shows a few strong spots, (nos. 21-24 inclusive), each of which contain a number of A residues and one U residue, judging by their relative positions on the fingerprint, and a number of weaker ones. Measurement of the radioactivity of each of these longer oligo-nucleotides show at least two major groupings. Each spot in the group of strong A-rich ones, nos. 21-24, occurs, on average, once every 120-150 bases and there is a second group, of 14 oligo-nucleotides, nos. 1,2,6,7,10,12,13,15,16,20,31,33,41 and 42, each of which occurs, on average, once every 400-600 bases. There is also a third group of oligonucleotides which comprises several minor components, each of which occurs about every 1,000 bases.

The pattern of this third group of spots is not regularily repeated in all the fingerprints. In some cases the number of spots

PLATE I

RNAase T_1 Fingerprint of ^{14}C-GTP labelled $cRNA_H$.

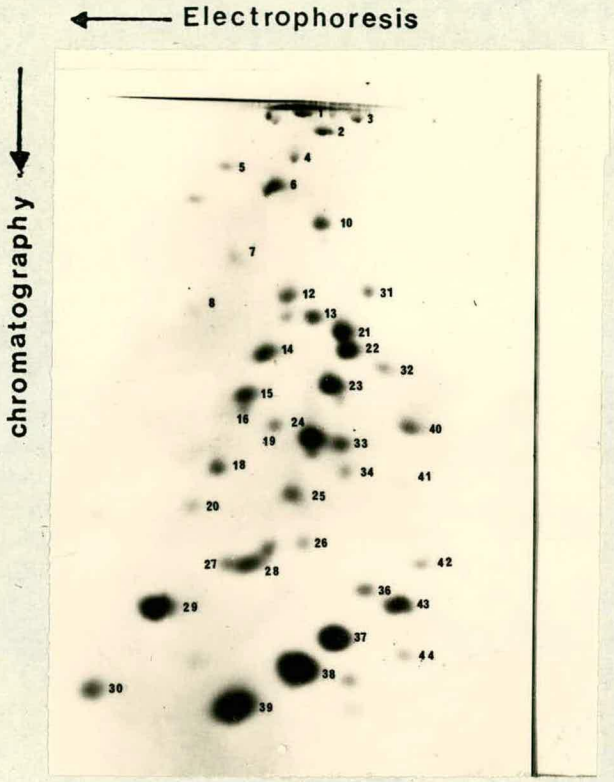
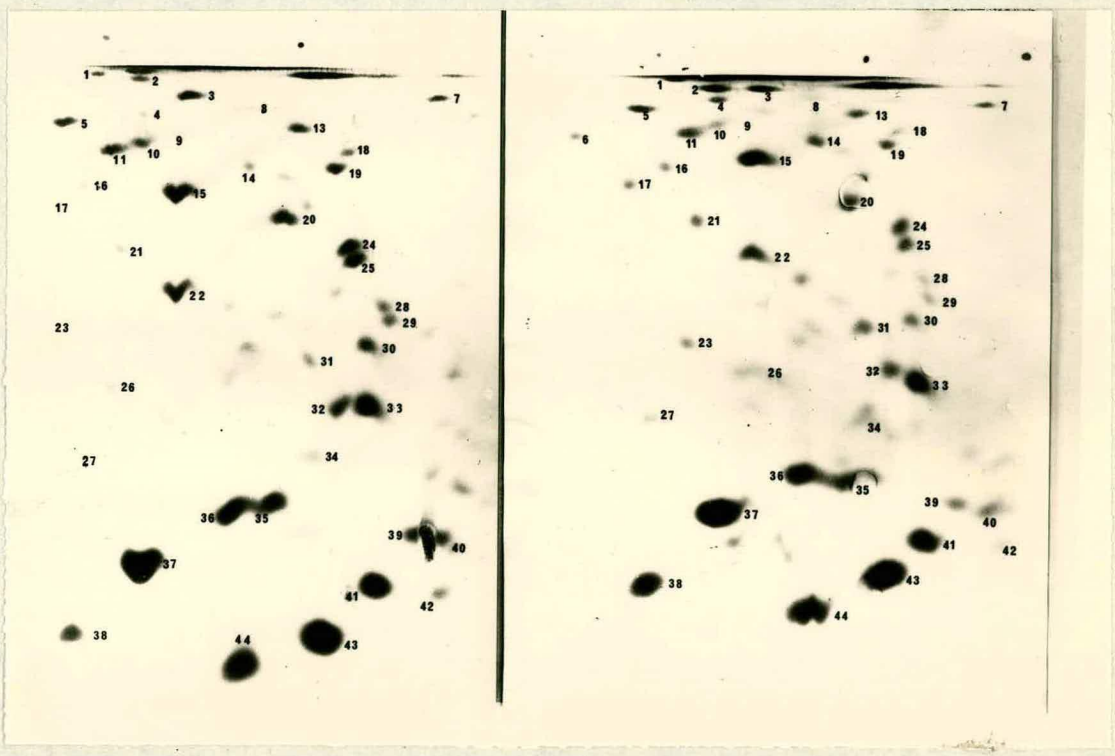


PLATE II

RNAase T_1 Fingerprints of ^{14}C-GTP labelled $cRNA_L$.



is reduced, whereas in others it is increased. These differences may reflect altered patterns of initiation and transcription and the spots themselves may arise from a more diverged group of sequences, such as may be located at the boundaries between satellite and non-satellite DNA, see section on 'the evolution of the sequence' in chapter IV. Furthermore, the polymerase may not transcribe all regions of the satellite with equal efficiency. It does not transcribe the light strand of the native satellite at all and may exhibit similar preferential initiation at certain sequences in the isolated strands.

In addition to these groups, there are a number of small oligonucleotides, mainly dimers and trimers, which occur at high frequency, (Table II).

The fingerprint of the light strand transcript, $cRNA_L$, shows a complementary pattern of sequences. The most prominent spots are rich in U residues and there at least two major groupings. The most prominent spots, nos. 1,2,3,5,15,20,22 and 24, occur, on average, once every 400-600 bases and there is a similar group of long U-rich oligonucleotides which occur at a lower frequency. There are also the small oligonucleotides which occur at high frequency. Although the $cRNA_L$ shows groupings of sequences, the grouping at 120-150 bases, seen in the analysis of the $cRNA_H$, is not found in this analysis. The reasons for this will be discussed in the following section.

The presence of groups of oligonucleotides, each one having the same repeat length, immediately suggests the presence of periodicities within the satellite, (chapter IV).

TABLE II

Quantitative Analysis of RNAase T1 Digestion Products of Mouse Satellite cRNA.

	Spot no.	Ave. cpm	% total cpm	Repeat Length
$cRNA_H$	1	251	1.18	423
	2	288	1.35	380
	3	174	0.89	609
	4	85	0.40	1,250
	5	57	0.31	1,600
	6	225	1.0	500
	7	45	0.2	2,500
	8	64	0.3	1,600
	9	51	0.2	2,500
	10	220	1.0	500
	11	66	0.3	1,600
	12	186	0.9	555
	13	78	0.37	1,350
	14	258	1.2	416
	15	218	1.0	500
	16	208	0.95	510
	17	79	0.37	1,350
	18	73	0.3	1,600
	19	97	0.5	1,000
	20	180	0.83	600
	21	702	3.3	151
	22	678	3.2	156
	23	695	3.3	151
	24	717	3.4	147

TABLE II (Cont.)

Spot no.	Ave. cpm	% total cpm	Repeat Length
25	193	0.9	555
26	128	0.61	790
27	480	2.3	220
28	399	1.9	250
29	1,340	8.0	60
30	362		
31	190	0.91	550
32	132	0.60	800
33	231	1.1	450
34	77	0.36	1,390
35	133	0.63	760
36	186	0.89	540
37	2,171	10.2	49
38	2,869	13.4	37
39	2,369	11.1	45
40	170	0.8	625
41	81	0.39	1,200
42	95	0.45	1,000
43	650	3.1	155
44	119	0.57	840
other, minor spots	750 est.		
origin	2,200		
Total	21,000		

TABLE II (Cont.)

	Spot no.	Ave. cpm	% total cpm	Repeat Length
$cRNA_L$	1	452	1.29	550
	2	613	1.75	410
	3	450	1.29	550
	4	248	0.71	1,000
	5	560	1.6	440
	6	258	0.70	1,000
	7	GTP		
	8	165	0.47	1,500
	9	173	0.49	1,500
	10	207	0.59	1,200
	11	335	0.95	750
	12	GDP		
	13	329	0.94	760
	14	260	0.73	1,200
	15	963	2.75	260
	16	125	0.35	2,000
	17	207	0.59	1,500
	18	308	0.88	810
	19	378	1.08	660
	20	515	1.47	485
	21	238	0.69	1,000
	22	455	1.3	550
	23	160	0.45	1,600
	24	465	1.32	550
	25	388	1.10	660

TABLE II (Cont.)

Spot no.	Ave. cpm	% total cpm	Repeat Length
26	127	0.36	2,000
27	121	0.32	2,000
28	263	0.75	950
29	270	0.78	920
30	277	0.80	900
31	336	0.96	740
32	347	0.99	850
33	704	2.0	350
34	137	0.39	2,000
35	805	2.3	310
36	840	2.4	300
37	3,570	10.2	70
38	1,785	5.1	140
39	241	0.70	1,000
40	273	0.78	900
41	1,190	3.4	210
42	126	0.35	2,000
43 44	4,165	11.9	60
other, minor spots	2,800 est		
origin	7,700		
Total	35,000		

TABLE II (Cont.)

Notes:-

Each spot was counted for 10 min. in a gas flow counter and averaged to give the counts per minute, (cpm), which were subsequently corrected for background, (-15 cpm for the $cRNA_H$ digestion products; 25 cpm for the $cRNA_L$ digestion products), to produce the final, Ave. cpm, figure shown in the table.

The Repeat Lengths of the $cRNA_L$ digestion products are slightly lower than those quoted in Biro et al. (1975), owing the omission of spots nos. 7 and 12 which were subsequently identified as GTP and GDP, respectively.

Repeat Lengths were calculated from the formula:-

$$\text{Repeat length} = \frac{100}{X \times P},$$

where 'X' is the percentage of the total radioactivity, obtained from RNAase T_1 digests of the appropriate cRNA, contained by each oligonucleotide and 'P' is the proportion of guanine in each cRNA transcript: 14% in the $cRNA_L$ and 22% in the $cRNA_H$, data taken from Walker, (1971a).

The repeat lengths of the oligonucleotides must be considered a maximum. Any contaminating enzyme activity, such as phosphatase or other nuclease, or over- or under-digestion of the cRNA, or cyclization of the products would reduce the amount of pure T_1 digestion products and increase the amounts of minor components which would lead to an increase in the apparent complexity of the satellite. The presence of more diverged sequences, discussed above, would also increase the observed repeat lengths of the T_1 digestion products and so would any small amounts of contaminating, non-satellite DNA. Such DNA might, for instance, have the same buoyant density as the satellite in the silver-containing gradients used to prepare the DNA and hence be co-purified with the satellite.

However, two lines of evidence suggest that the satellite is, in fact, pure. Firstly, the purity of a sample of satellite prepared by this method was confirmed by analytical ultracentrifug--ation and, secondly, alkaline CsCl centrifugation produced two equal-sized peaks for the heavy and light strands of the satellite. Any contaminating main-band DNA would have been observed as an increased heavy strand peak under such conditions.

Fingerprint Analysis of the cRNA.

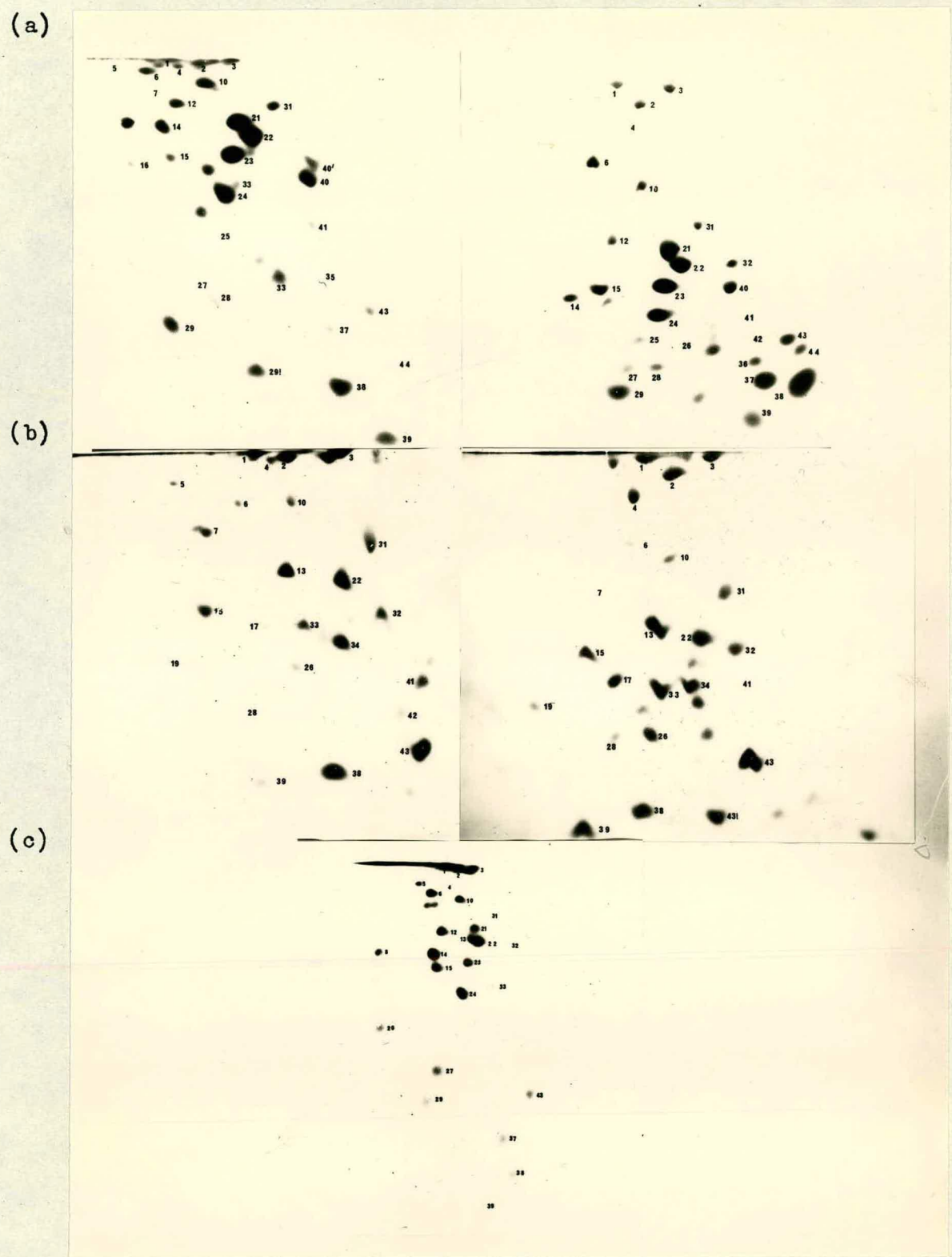
(a) Heavy Strand

Preliminary analysis of transcripts of the native satellite made using α^{32}P-GTP and α^{32}P-UTP had been obtained previously by A.Carr-Brown and the information obtained is summarized in Table 3.

Further analysis of the cRNA was subsequently performed. Transcripts made with α^{32}P -UTP, -CTP and -ATP as label were analysed by RNAase T_1 digestion and transcripts made with α^{32}P -ATP, -CTP and -GTP as label were analysed by RNAase A digestion.

PLATE III

RNAase T_1 Fingerprints of ^{32}P-labelled $cRNA_H$.



(a) labelled with ^{32}P-ATP

(b) labelled with ^{32}P-CTP

(c) labelled with ^{32}P-UTP

TABLE III

Nucleotides from RNAase T_1 Digest of $cRNA_H$.

Spot no.	Summary of G and U labels.	U Panc.	C Panc.	C Base Comp.	A Panc.	A Base Comp.	R.L.	Composition	Sequence
1	-AG	AC	A,C	A_3U AC C	A,U	C	423	$(A_3U)_2(ACC)(AC)$-AG	
2	-CUG	AU	U	A_7U AU	A,U		380	$(AC)(A_7U)(A_3C)(AU)$-CUG	
		AG	A	A_3C	C				
3	-G	A_3U	U	A_4U A_3U	A,U		609	$(A_3UC)(A_4U)(CC)$-G	
		C	C		C				
4		A_4C C	A C	A_4U	A		1,250	$(A_4C)(CC)(A_4U)$-G	
5	-AAUG -AU-			AAU AU	A,U		1,600	$(A_5U_4)G$	AUAUAUAAUG
6	-AUG -AG			A_4U U	A,U		500	$(A_5U_3)G$	UA_4UUAG UA_4UAUG
7	-AAUG -CUUG			AAU	A,U		600	$(A_3U_3C)G$	AAUACUUG ACUUAAUG
8	-AUG						1,600	$(A_3U_3)G$	
9	-AUAUUG						2,500	$(A_2U_3)G$	AUAUUG

TABLE III (cont.)

Spot no.	Summary of G and U labels.	U Panc.	C Panc.	C Base Comp.	A Panc.	A Base Comp.	R.L.	Composition	Sequence
10	-AUG(G)	A_4U A_3U AC		U, A	A_4U A_3U AC C	A C	500	$(A_5U_2C_2)G$	$CA_4UCAUG(G)$ $ACA_3UCAUG(G)$
11	-UG						1,600	$(C_2A_5U)G$	
12	-AUG(G)				AU A_3U	A, U, G	550	$(A_3U_2)G$	$A_3UAUG(G, A)$
13		AAU AU AC		U, A C			1,350	$(A_3U_2C_3)G$	AAUCCACUG AUCCACAUG
14	-AUG(G)	AAU AU			AAU AU	A	416	$(A_3U_2)G$	AAUAUG(G)
15	-CUG	AU AC	AC	A	AU G	U G	500	$(A_2U_2)G$	AUACUG(A)
16	-AUG(G)				AU	U	510	$(A_2U_2)G$	AUAUG(G)
17	-AUG -CUG -UUG		AC	A, C			1,350	$(A_2U_2C_2)G$	CACUAUG ACCUAUG UACACUG ACACUUG
18	-AUUG(G) -UAUG(G)						1,600	$(AU_2)G$	AUUG(G) UAUG(G)
19	-CUUG(G) -AUG(G)		AC	A			1,000	$(ACU_2)G$	ACUUG(G) UCAUG(G)

TABLE III (cont.)

Spot no.	Summary of G and U labels.	U Panc.	C Panc.	C Base Comp.	A Panc.	A Base Comp.	R.L.	Composition	Sequence
20	-UUG(G)	U					600	UUG	UUG(G)
21	$-A_5UG$	A_5U			A_5U	A,G	150		$A_5UG(A)$
22	-CUG	A_4C	A_4C	A	A_4C	A,G	156		$A_4CUG(A)$
23	$-A_4UG$	A_4U			A_4U	A,G	151		$A_4UG(A)$
24	$-A_3UG$	A_3U			A_3U	A	147		A_3UG
25	-AAUG				AAU	A,U	555	$(A_2U)G$	AAUG
	-AUAG				AU				AUAG
					AAG				UAAG
26	(ACU)G	AC	AC	A	C	C	790	(ACU)G	ACUG
			AU	U					AUCG
									CAUG
27	AUG	AU			G	G	220		AUG
28	UAG		AG	G	U,G	U,G	250		UAG(A,C)
29	UG				G	G	60		UG(A)
30	UG				G	G			UG(A)!
31	-ACG(G)		A_4U	U	A_4U	A	550	$(A_5UC_2)G$	$A_4UCACG(G)$
	-UCG(G)		AC	A	AC	C			$ACA_4UCG(G)$
					C				
32	-CUG		A_4C	A,C	A_4C	A,C	800	$(A_4C_3U)G$	CA_4CCUG
					A_3C				CA_3CACUG
			AC		AC				$CACA_3CUG$
			C		C				

TABLE III (Cont.)

Spot no.	Summary of G and U labels.	U Panc.	C Panc.	C Base Comp.	A Panc.	A Base Comp.	R.L.	Composition	Sequence
33	-UG		A_3C	A	A_3C	A,C	450	$(A_3CU)G$	$CA_3UG(C)$
					A_3U				$A_3CUG(C)$
			G	G					
34	-AUG(G)		AC	A,C	AC		1,390	$(A_2UC_2)G$	ACCAUG(G)
	-CUG(G)		C	C	C	C			CACAUG(G)
	-ACG(G)								ACACUG(G)
35					A_3G	A	760		A_3G
36	-AAG				AAG	A	540		AAG
37	AG	AG			AG	G	49		AG
38	G	G	G	G	G	G	20	G	G
39	G!	G	G	G	G	G		G!	
40	-AG(G)				A_5G	A	625		$A_5G(G)$
41	$-A_3CG(G)$		A_3C	A	A_3C	A,C	1,200	$(A_3C)G$	AAACG(G)
			A_2C		A_2C				AACAG(G)
42	-CAAG				AAG	A,G	1,000	$(A_2C)G$	CAAG(A)
	-ACAG				AC	C			ACAG(A)
					AG				
					C				
43	-ACG		AC	A	G	G	155		ACG(A)
44	-CG		G	G	G	G	840		CG

The major sequence isomer in a mixture is underlined.

TABLE III (cont.)

Notes

Cyclic nucleotides are indicated by:- !

'Panc.' refers to the products obtained from each oligonucleotide after secondary digestion with Pancreatic RNAase, (RNAase A).

'Base comp.' refers to the products obtained from each oligo--nucleotide after alkaline hydrolysis.

Data from ^{32}P-CTP labelled cRNA is based on two separate analyses.

Data from ^{32}P-ATP labelled cRNA is based on three analyses.

Data from ^{32}P-UTP labelled cRNA is based on one analysis.

The spot numbering system does not necessarily correspond to that used in Biro et al. (1975).

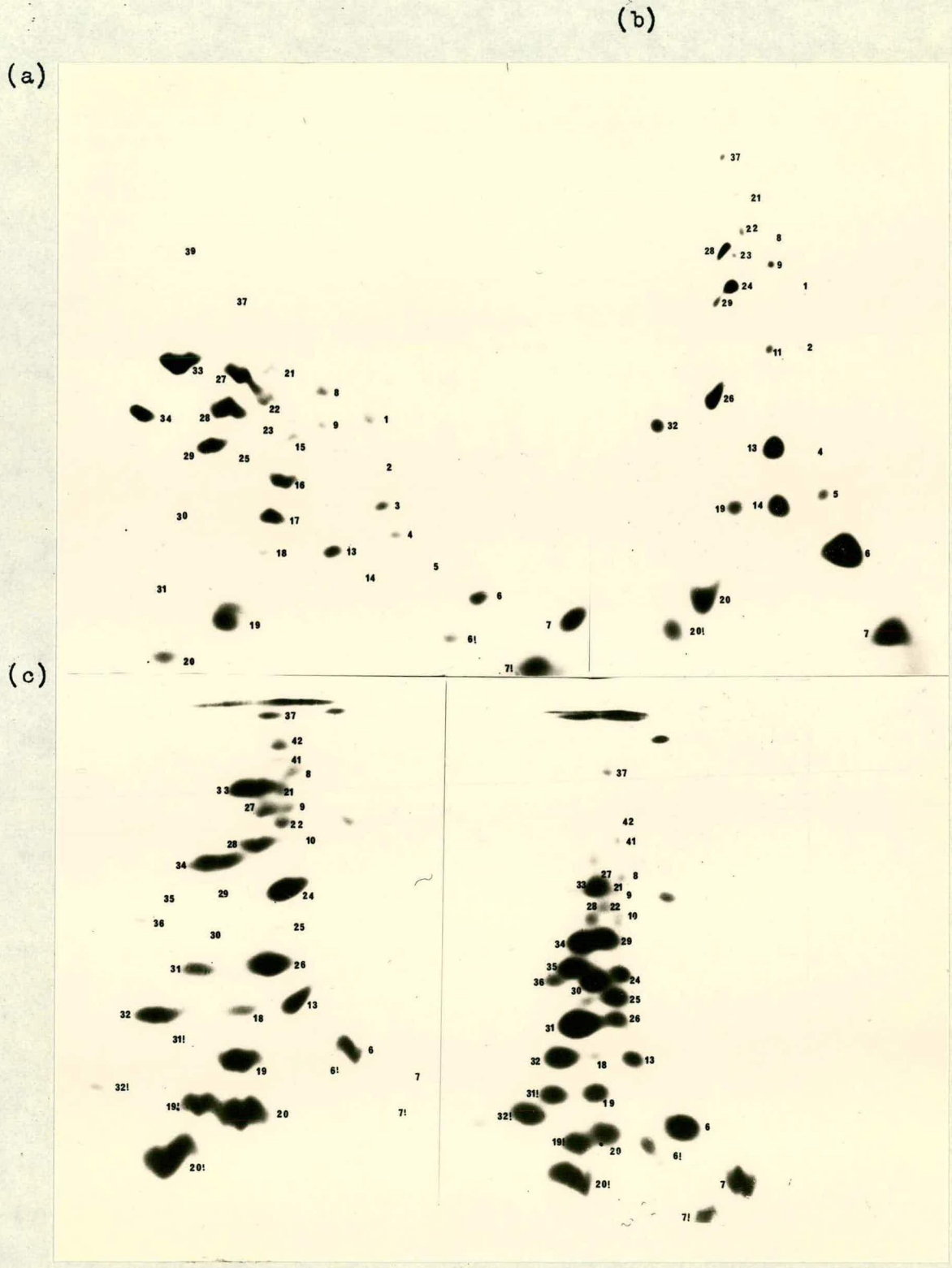
The sequences of most of the oligonucleotides could be deduced by simple secondary analysis. In each case, the spots were cut out, counted, and the radioactivity eluted with 30% triethyl ammonium carbonate. One portion of each spot was digested with the reciprocal RNAase, for example RNAase T_1 digestion products were digested with RNAase A, and the products of digestion separated and identified by paper ionophoresis at pH 3.5 and also in 7% formic acid. The second portion was subjected to nearest neighbour analysis by alkaline hydrolysis and the bases separated by paper ionophoresis at pH 3.5, (Sanger and Brownlee, 1967; Adams et al. 1969). The results are shown below, Plates III and IV and Tables III and V.

Some of the larger oligonucleotides were only partially sequenced and, for these, the composition of the secondary digests are given. Such long oligonucleotides produced complex spots and were composed either of more than one sequence or more than one isomer of the same sequence, and, in addition, were not found in all the fingerprints. Consequently analysis of these spot proved difficult and unequivocal sequences cannot always be given. Where a complex spot has more than one sequence isomer the major component is underlined, Tables III and V.

The sequences of the four major oligonucleotides, referred to above, spots nos. 21-24, are:- A_5UG, A_4CUG, A_4UG and A_3UG respectively. These sequences are all clearly related - the last three can each be derived from the first by only one base change. Nearest neighbour analysis shows a further relationship, namely that the base adjacent to the 3' G residue is 'A', in the case of spots nos. 21, 22 and 23. All four sequences show the expected complementarity to the group of T_nC tracts observed in the pyrimidine tract

PLATE IV

RNAase A Fingerprints of ^{32}P-labelled cRNA$_H$.



(a) labelled with ^{32}P-ATP

(b) labelled with ^{32}P-CTP

(c) labelled with ^{32}P-GTP

analysis of the heavy DNA strand. For example, there is a pyrimidine tract, T_5C, which has a repeat length of 130 b.p. and which corresponds to the (G)A_5-UG sequence, seen as spot no. 21, (Table V).

In addition to these major oligonucleotides there are several of lesser importance which contain A_nU sequences such as A_3U and A_4U, for example spots nos. 1,2,3 and 6, or which contain sequences very closely related to A_3U and A_4U, for example spots nos. 5,8,15,16,40 and 41. These spots could have arisen from an original A_4U-containing sequence by only a few base changes. For example, spot no. 10, (G)A_4UCAUG could have been derived from a basic (G)A_4UGA(UG) sequence by one G to C base substitution. Spot no. 14, (G)AAUAUG(G), could have arisen from (G)A_4UGA by an A to U and an A to G change. Similarily, spot no. 12, (G)A_3UAUG(A) from (G)A_5UG(A) by one A to U change. Nearly all of these secondary oligonucleotides, which have a repeat length of between 400-600, could have been formed from one of the major ones by divergence.

The RNAase A fingerprints of the $cRNA_H$, Plate V, also show a major group of A-rich spots, many of which contain one, (spots nos. 8,9,27,28 and 29), or two, (spots nos. 21,22,33 and 34) G residues at their 5' termini and, similarily, a C residue (spots nos. 8,9,21 and 22) or a U residue, (spots nos. 27,28,29,33 and 34) at their 3' ends. These spots correspond to the (G)A_nUG and (G)A_nCUG sequences produced in the RNAase T_1 digests.

Since RNAase A cleaves after two bases, C and U, detailed quantitative analysis of RNAase A fingerprints would not be as informative or as straightforward as analysis of RNAase T_1 fingerprints, although the intensity of each spot as visualized on the radioautograph does give an indication of its relative

TABLE IV

Nucleotides from RNAase A digests of $cRNA_H$ and $cRNA_L$.

Spot no.	Composition	Sequence	
		$cRNA_H$	$cRNA_L$
1	A_6C	AAAAAAC	-
2	A_5C	AAAAAC	-
3	A_4C	AAAAC	-
4	A_3C	AAAC	-
5	A_2C	AAC	AAC
6	AC	AC	AC
6a	AC!	AC!	AC!
7	C	C	C
7a	C!	C!	C!
8	$(GA_6)C$	AAGAAAAC	-
9	$(GA_5)C$	GAAAAAC	-
10	$(GA_4)C$	GAAAAC	-
11	$(GA_3)C$	AGAAC, AAGAC	-
12	$(GA_2)C$	x	-
13	(GA)C	GAC	GAC, AGC
14	GC	GC(C)	GC

TABLE IV (Cont.)

Spot no.	Composition	Sequence	
		$cRNA_H$	$cRNA_L$
15	A_5U	AAAAAU	-
16	A_4U	AAAAU	-
17	A_3U	AAAU(A)	-
18	A_2U	AAU	-
19	AU	AU	AU
20	U	U	U
20a	U!	U!	U!
21	$(G_2A_5)C$	GAGAAAAC AGGAAAAC	-
22	$(G_2A_4)C$	GAGAAAC(A)	-
23	$(G_2A_3)C$	GGAAAC	-
24	$(G_2A_2)C$	GAGAC	-
25	$(G_2A)C$	GAGC, GGAC	GGAC, GAGC AGGC
26	GGC	GGC	GGC
27	$(GA_5)U$	GAAAAAU, AAGAAAU	-
28	$(GA_4)U$	GAAAAU, AGAAAU	-
29	$(GA_3)U$	GAAAU	-

TABLE IV (Cont.)

Spot no.	Composition	Sequence	
		$cRNA_H$	$cRNA_L$
30	$(GA_2)U$	GAAU	x
31	(GA)U	GAU	x
32	GU	GU	-
33	$(G_2A_4)U$	GGAAAAU, GAGAAAU	-
34	$(G_2A_2)U$	GGAU	-
35	$(G_2A)U$	x	-
35a	GGU	GGU	GGU
37	x	x	-
38	x	x	-
39	x	x	-
40	GGGU	GGGU	-
41	$(G_2A_5)U$	-	x
42	$(G_3A_4)U$	-	x

Notes

(-) indicates nucleotide is absent or present in only trace amounts.

(x) indicates composition of oligonucleotide is unknown.

frequency in the sequence.

The majority of these A-rich sequences are of the form:-

$$\begin{matrix} G- \\ GG- \\ GAG- \end{matrix} A_n - \begin{matrix} U \\ C \end{matrix}$$

and, by combining this information with that obtained from the RNAase T_1 digests, (and the pyrimidine tract analysis), a possible longer sequence for each major spot can be derived, (Table V).

For example, spot no. 21 of the T_1 analysis, (G)A_5UG(A), may be part of a longer sequence:- U/C-GA_5UGA as A_5U occurs predominantly as GA_5U, rather than, for example, GGA_5U or GAGA_5U, in the RNAase A digests. Similarily, the principal T_5-containing tract in the pyrimidine tract analysis occurs as T_5C which corresponds to U/C-GA_5-U/C in the $cRNA_H$ and also supports a possible U/C-GA_5UGA sequence for spot no. 21.

The major T_4-containing pyrimidine tracts are T_4C, T_4CC, and T_4CTC and T_4CTT, which correspond to:- GA_4, GGA_4, GAGA_4, and AAGA_4 respectively. The principal A_4U sequences seen in the RNAase A digests are:- GGA_4U and, toa lesser extent, GA_4U. Consequently, spot no. 22 of the RNAase T_1 analysis, (G)A_4UG(A), may be part of a longer sequence:- U/C-GGA_4UGA. The main A_4C sequences found from the RNAase A digests are as:- GAGA_4C and AAGA_4C, hence spot no. 23, (G)A_4CUG(A), may occur as:- U/C-GAGA_4CUGA and, also, as:-U/C-AAGA_4CUGA, (Table V).

However, such evidence for possible longer sequences is, of necessity, rather indirect and only a direct method of sequencing longer stretches of the satellite can provide definitive evidence for such proposed longer sequences.

TABLE V

Sequences of Principal Oligonucleotides Found in the Pyrimidine Tract Analysis and Satellite $cRNA_H$.

Spot no.	Sequence from RNAase T_1 digest	Sequence from RNAase A digest	Sequence from P.T. analysis	Longer Sequence
24	$(G)A_3UG$	GA_3U	GA_3	U/C-GA_3UG
		$GAGA_3U$	$GAGA_3$	-$GAGA_3UG$
23	$(G)A_4UG(A)$	GA_4U	GA_4	-GA_4UGA
		GGA_4U	GGA_4	-GGA_4UGA
			AGA_4	-$GAGA_4UGA$
22	$(G)A_4CUG(A)$	$GAGA_4C$	$GAGA_4$	-$GAGA_4CUGA$
		$AAGA_4C$	$AAGA_4$	-$AAGA_4CUGA$
		$AGGA_4C$		-$AGGA_4C$
		GA_4C		-GA_4C
21	$(G)A_5UG(A)$	GA_5U	GA_5	-GA_5UGA

(b) Light Strand

Transcripts of the purified light strand of the satellite DNA were made using α^{32}P-ATP, α^{32}P-CTP and α^{32}P-GTP as label and analysed in the same way as the various heavy strand transcripts. The information obtained from these transcripts complemented and extended that obtained from the heavy strand, (Table VI).

The major feature of the $cRNA_L$ is the large number of U-rich spots found in the RNAase T_1 fingerprints, (Plate V). Several of the most frequent of these U-rich spots contain 5,6 or 7 U residues, for example, spots nos. 2,3,5,9,10 and 11, (judged by their relative positions on the fingerprint). These tracts show the expected complementarity to the A_nUG and A_nCUG sequences found in the $cRNA_H$ and contain a run of U residues terminating in a C residue. In general, this C residue may be followed by:- -AG, -ACAG or -CACAG and these would correspond to $-CUGA_n$, $-CUGUGA_n$ and $-CUGUGGA_n$ sequences at the 5' ends of the A_nUG and A_nCUG sequences found in the $cRNA_H$.

This pattern can be illustrated by analysis of spots nos. 9, 10 and 11 of the $cRNA_L$, which differ from each other by the addition of one C residue. Spot no. 9 contains three C residues, spot no. 10 contains two C residues and spot no. 11, the major one, has one C residue. The sequence at the 3' end of spot no. 11 is $-U_nCAG$, where n is probably 5, and that at the end of spot no. 10 is $-U_nCCAG$ and $-U_nCCG$, and that at the end of spot no. 9 is $-U_nCCACG$. These three termini would correspond to $-CUGA_n\ldots$, $-CUGGA_n\ldots$ and $-CGGA_n\ldots$, and $-CGUGGA_n\ldots$ in the $cRNA_H$, all of which are possible sequences for the 5' end of some of the GA_nUG and GA_nCUG tracts. The principal sequence in this family, spot no. 11, $-U_nCAG$, which

PLATE V

RNAase T_1 Fingerprints of ^{32}P-labelled $cRNA_L$.

labelled with:-

(a) ^{32}P-ATP

(b) ^{32}P-CTP

(c) ^{32}P-GTP

TABLE VI

Nucleotides from RNAase T_1 Digest of $cRNA_L$.

Spot no.	C Panc.	C Base comp.	A Panc.	A Base comp.	G Panc.	G Base comp.	R.L.	Sequence
1			C	C	U C	U C	550	-CA-...-UG -CG
2	U	U	AAU C	A C	AU AG	U A	410	-UCA-...-AUG -AG
3	U	U	C AAU	C A	AAU	U	550	$-U_5CAAUG$
5	AG U	G U	C	C	AG	A	440	$-U_6CAG(C)$ $U_5CUCAG(C)$ $UCU_5CAG(C)$
7	CTP		ATP		GTP			
9	C U AC	C U A	C	C	AC	C	1,500	$-U_5CCACG$
10	C U	C U	C	C	AG C	A C	1,200	$-U_5CCAG$ $-AU_5CCG$
11	U	U	C	C	AG	A	750	$-U_5CAG$
12	CDP		ADP		GDP			
13	AC	A	C	C	AG	A	760	$-U_4CCACAG$ $-U_4CACCAG$
15	U	U	C	C	AG	A	260	$-U_4CAG$
18	AC C U	A C U	AC C	C C	AC	C	810	$(A_2U_3C_2)ACG$

TABLE VI (Cont.)

Spot no.	C Panc.	C Base comp.	A Panc.	A Base comp.	G Panc.	G Base comp.	R.L.	Sequence
19	C U	C U	C	C	AG	A	660	$(A_2U_3C_2)$AG eg. ACAU$_3$CCAG or AU$_3$CCACAG
20	C U	C U	C	C	AG	A	485	(U_3C_2A)AG eg. AU$_3$CCAG
22	U	U	(C)	(C)	C	C	550	AUUUCG CAUUUG
23	U	U			C	C	1,600	UUUCG
32			AAU	A	AAU	U	850	AAUG
33	U	U	AAG C	A C	AAG	A	350	UCAAG
35	AU	U			C	C	310	AUCG
36					AU	U	300	AUG
37			(G)	(G)	U	U	70	UG(A)
38	(G)	(G)			U	U	140	UG(C)!
39			AAG	A	AAG	A	1,000	AAG
40	AC	A	C	C	AC AG	C A	900	ACG CAG

TABLE VI (Cont.)

Spot no.	C Panc.	C Base comp.	A Panc.	A Base comp.	G Panc.	G Base comp.	R.L.	Sequence
41					AG	A	210	AG
42					C	C	2,000	CG
43	G	G			G	G	60	G
44			G!	G!				

Notes

Digestion products present in only small quantities are shown in brackets, for example:- (G).

Spot numbers omitted refer to oligonucleotides present only in trace amounts.

The spot numbering system does not necessarily correspond to that used in Biro et al. (1975).

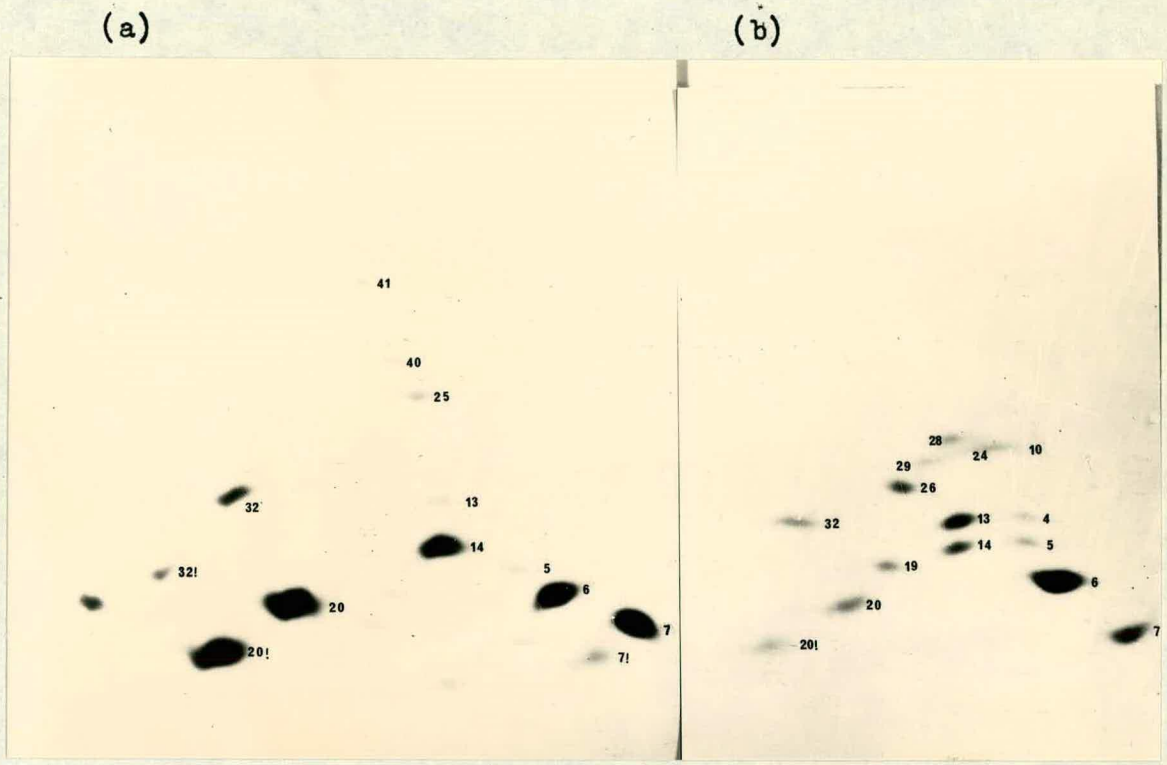
corresponds to $CUGA_n$ in the $cRNA_H$, and is most likely to be the sequence at the 5' end of some of the U/CGA_5UGA sequences and would extend it to:- ...$CUGA_5UGA$... . Such a sequence would be in keeping with the possible longer sequences shown in Table V. Other relationships between the $cRNA_L$ analysis and the nature of the A-rich tracts will be discussed in the following chapter.

The quantitative features of the $cRNA_H$ and the $cRNA_L$, referred to earlier, appear to be quite different. The $cRNA_L$ does not show a grouping of oligonucleotides having the repeat length of 120, found for the $cRNA_H$. In addition, RNAase T_1 fingerprints of this cRNA contain a greater number of U-rich spots than complementary, A-rich spots occur in the RNAase T_1 fingerprints of the $cRNA_H$, which contains a higher proportion of G residues (20% as opposed to 14%) and would, therefore, be expected to produce more, not less, spots after RNAase T_1 digestion. This apparent contradiction can be explained by the effects of divergence. The explanation of the lower yield and greater total number of spots is that the fewer the number of G residues, the greater the distance between them. Hence there can be more instances of divergence between G residues and, therefore, more spots will be produced after RNAase T_1 digestion. The spots produced will be closely related, for example spots nos. 9, 10 and 11 differ only by their number of C residues and so do nos. 18 and 19 and also nos. 23 and 24. Furthermore, the prominent U-rich spots from the $cRNA_L$ are, on average, longer by about 2-3 b.p., than the prominent A-rich spots from the $cRNA_H$.

Very little information can be obtained from the RNAase A fingerprints of the $cRNA_L$, (Plate VI and Table IV). The U-rich tracts are completely degraded by the enzyme and no purine-rich

PLATE VI

RNAase A Fingerprints of ^{32}P-labelled cRNA$_L$.



(a) labelled with ^{32}P-GTP

(b) labelled with ^{32}P-CTP

sequences are present in this cRNA, consequently, no large spots are found in the fingerprints. The most prominent sequences are:- U, GU, C, AC and GC. This is very much as expected on the basis of the results obtained from previous fingerprints and from the pyrimidine tract analysis.

The satellite DNA has, therefore, been shown to consist of related sequences which can be grouped into families, according to to the relative frequences of the RNAase T_1 digestion products of the cRNAs transcribed from the heavy and light strands of the DNA. In the following chapter the arrangement of these sequences in the DNA will be examined.

CHAPTER THREE

Sequence Analysis of Mouse Satellite DNA

Introduction.

Until recently, DNA sequencing has been much more difficult than RNA sequencing owing to the lack of suitable site-specific endonucleases with which to cleave the DNA. DNAases equivalent to RNAases such as T_1, A and U_2 are not known. Bernardi et al. (1974) have investigated the use of DNAases as possible sequencing tools, but the results show no useful specificities. Sanger and co-workers have made use of Endonuclease IV, (isolated by Sadowski and Hurwitz, 1969), which can cleave DNA between TpC pairs, to sequence fragments of phage ØX 174, (Robertson et al. 1973, Ziff et al. 1973, and Galibert et al. 1974). However, apart from this enzyme, specific degredation of DNA has, in general, been limited to pyrimidine tract analysis.

However, a number of discoveries have recently been made which greatly facilitate DNA sequencing. Firstly, the discovery of type II restriction enzymes, (Smith and Wilcox, 1970; Kelly and Smith, 1970) has provided a method of producing DNA sequences of different lengths with defined 5' and 3' ends. A variety of methods can now be used in conjunction with restriction fragments to sequence the DNA, (see reviews by Salser, 1974; Murray and Old, 1974; Murray, 1974, Sanger, 1975) and new methods are being devised, (Sanger and Coulson, 1975; Gilbert and Maxam, unpublished results).

One particularily useful approach involves sequence-specific hybridization of fragments of DNA, (or RNA), to isolated DNA strands. The fragments may either be synthetic oligonucleotides or

restriction fragments which are used as primers for extension by a polymerase to produce a copy of the isolated DNA strand. The newly synthesized DNA can be analysed by techniques such as ribosubstitution or by the 'plus-and-minus' method of Sanger and Coulson, (1975), used in this work.

Berg et al. (1963) discovered that it is possible to artificially alter the properties of the E.coli DNA polymerase I enzyme, by substituting Mn^{++} ions for Mg^{++} ions in the incubation mix, such that it will incorporate the corresponding ribonucleotide, as well as deoxyribonucleotides, into the nascent DNA chain. cDNA can be synthesized containing a single ribonucleotide which is then sensitive to cleavage either by alkaline hydrolysis or by a ribonuclease such as RNAase T_1. Ribosubstitution has been used by a number of groups in sequencing DNA, for example, Van der Sande et al.(1972), Salser et al. (1972) on M13 phage, Sanger et al. (1973) on a phage f1 fragment, Maniatis et al. (1974) on a repressor binding site in phage Lambda and Fry et al.(1972) who used both ribosubstituted cDNA and conventional cRNA to sequence the HS-β satellite of the kangaroo rat, Dipodomys ordii, and obtained identical sequences by each method, thus confirming the reliability of the techniques, (Whitcombe et al. 1974).

The availability of isolated single strands of mouse satellite DNA and Eco.RII restriction fragments immediately suggested an approach for the direct sequencing of a specific length of the satellite DNA, from which it was hoped to be able to order the A-rich tracts and sequence the intervening regions. The method of Sanger and Coulson, (1975), was adopted as it would enable a long section of the satellite to be examined in a single experiment, even if the resulting sequence was not complete, owing both to

the limitations of the technique and to the divergence of the satellite.

Southern, (1975), had calculated the overall divergence in the reassociation register of the satellite to be 3%. It was not known whether this would affect the sequencing method which depends on the separation of synthesised stretches of polynucleotides up to 100 bases in length and which differ in length by only one base. Although homogeneous polynucleotides give sharp bands on the polyacrylamide gels used, it seemed possible that the introduction of one or two base changes in the 100 or so bases, (owing to sequence heterogeneity) could affect the separation. Furthermore, it was not known whether the divergence was exclusively due to base substitutions or whether it involved additions or deletions as well. The data of Southern, (1975), suggested that the monomer may have a maximum of 10% length heterogeneity, as measured by its band width on an agarose gel. However, the average length heterogeneity is not known, only the maximum heterogeneity.

In order to test whether all satellite sequences are found in each class of restriction fragment, a sample of satellite DNA was restricted to produce a terminal digest and a selection of such fragments was eluted, purified and subsequently transcribed and fingerprinted.

Materials and Methods

Separated, purified single strands of mouse satellite DNA were prepared on alkaline CsCl gradients, as described in Chapter Two.

Eco.RII restriction endonuclease was made by the method of Yoshimori, (1971) from *E.coli* fi^- R factor, (a gift of S.G. Hughes) and was found to be free of exonuclease activity.

Preparation of Satellite Restriction Digest

Incubations were carried out in 50mM NaCl, 6mM $MgCl_2$, 5mM II-mercaptoethanol, 10mM Tris-HCl, pH 7.4, for 4hr at 37°C. 4µl of enzyme in 10µl of incubation mix was required for each µg of satellite DNA. Reactions were terminated by extracting the enzyme with an equal volume of buffer-saturated redistilled phenol, removing the phenol by three extractions with ether and desalting the DNA on a Sephadex G75 column which was eluted with 1mM Tris-HCl, 0.1mM EDTA,(pH 7.4).

For the preparation of purified fragments for transcription, 30µg of satellite was digested with 75µl of enzyme for 18hr and the products separated on a 1.5% cylindrical agarose gel; 1cm in diameter, 20cm long. The sample was loaded in 5% glycerol, containing bromophenol blue as marker dye, and electrophoresis was carried out in E buffer, (Loening, 1967), for 18hr at 10mA, until the bromophenol blue had run about 14cm. The gel was soaked in E buffer containing 0.5µg/ml ethidium bromide for 30min and the DNA visualized by illumination under ultraviolet light at 254nm. Individual restriction fragments were recovered by slicing the gel with a flamed blade and dissolving the agarose either in 1 vol. of 5M sodium perchlorate at 60°C, diluting the solution to 2.5M perchlorate, 0.05M sodium phosphate, (pH 6.8), and adding a small amount of hydroxylapatite at 60°C to absorb the DNA, or by dissolving the gel in 1 vol. of saturated potassium iodide solution containing 0.05M sodium phosphate buffer and similarily absorbing the DNA onto hydroxylapatite at room temperature.

In both cases, the hydroxylapatite was pelleted, washed extensively with 0.05M sodium phosphate and subsequently with

water and the DNA eluted at room temperature with 0.5M phosphate buffer. About 0.5ml of buffer was usually sufficient and the phosphate was removed by dialysis against 1mM Tris-HCl, 0.1mM EDTA, (pH 7.5), and the DNA concentrated by evaporation under partial vacuum.

Transcription of Restriction Fragments.

Purified restriction fragments were transcribed in the usual way and the cRNA was labelled with α-^{32}P-GTP, (Radiochemical Centre, Amersham, 100Ci/mMol), and fingerprints made of RNAase T_1 digests.

Sequencing of the DNA.

The method used was essentially that of Sanger and Coulson, (1975).

Samples of an Eco.RII total restriction digest of mouse satellite DNA were hybridized to separated single strands of the satellite in the ratio 10µg single strand DNA : 1µg restriction fragments. Usually 5µg single strand DNA was used and hybridization was carried out in 50µl of restriction buffer for up to 4hr at 67°C. It was found necessary to desalt the DNA at this stage by chromatography on a Sephadex G75 column as otherwise yields of cDNA were often very small. Elution was with 1mM Tris-HCl, 0.1mM EDTA, (pH 7.4).

After concentrating the DNA by evaporation under partial vacuum, cDNA synthesis was carried out in 50µl of restriction buffer containing 0.05mM of each non-labelled deoxyribonucleotide and 5µCi of one or more α-^{32}P-labelled deoxyribonucleotide, (100Ci/mMol, the dATP, dCTP, dGTP and dTTP used were trial samples generously donated by the Radiochemical Centre, Amersham). In some experiments

all four deoxyribonucleotides carried the ^{32}P label. 10µl of E.coli DNA polymerase I, large fragment A, (Boehringer Corporation), was then added to the mixture and incubation carried out at 0°C. 10µl samples were removed at intervals, up to a maximum incubation time of 10min and the reaction terminated by addition to 10µl of 200mM EDTA. The cDNA was purified by chromatography on a Sephadex G75 column which was eluted with 1mM Tris-HCl, pH 7.4, and concentrated by evaporation under partial vacuum.

The minus incubation mixtures contained 0.25µl of the three dNTPs, 0.3mM each, 0.25µl of DNA polymerase I, one-eighth of the cDNA in a final incubation volume of 5µl. Incubation was at 0°C for 30min, in restriction-enzyme buffer.

The plus mixtures contained 2µl of the appropriate dNTP, 1mM, 0.25µl of T4 DNA polymerase, (a gift of S.G.Hughes), and one-eighth of the cDNA in a volume of 5µl of restriction-enzyme buffer. Incubation was for 30min at 37°C.

At the end of 30min, 2µl of Eco.RII enzyme was added to each incubation mix and restriction of the cDNA carried out for 30min at 37°C. Reactions were terminated by the addition of 25µl of deionized formamide containing 2µl of 200mM EDTA and xylene cyanol FF and bromophenol blue as marker dyes.

The solutions were heated at 100°C for 5min and layered onto a 12% polyacrylamide gel, (20cm x 40cm x 0.1cm) which had been made up in Tris-glycine buffer containing 8M Urea, (3.028g Tris, 14.4g glycine/L, pH about 8.0), (Peacock and Dingman, 1967; Sanger and Coulson, 1975). The buffer in the anode and cathode compartments was the same but lacked the urea. Electrophoresis was continued until the bromophenol blue had run off the gel, usually about 800v for 8hr, although voltages up to 1,700v were used. The gel

was run in a fume cupboard and cooled, when necessary, by the extractor fan. To avoid distortion of the bands due to localized heating effects, very thin glass plates were used to house the gel, about 1.5mm thick. After electrophoresis, the gel was removed from the glass plates and covered with polythene and radioautographed. It was found that the bands remained sharp for up to 4 days, but if longer exposures were required, the gel first had to fixed by soaking it in 10% acetic acid for 20min which, although it caused the gel to expand slightly, did not distort the pattern of the bands. Kodak 'Blue Brand' X-ray film was used for radioauto--graphy, although Kodak Royal RP-X-OMAT film was found to give a lower background.

Results

Analysis of cRNA transcribed from restriction fragments.

Southern, (1975) showed that mouse satellite DNA could be cleaved by Eco.RII restriction enzyme to give fragments which form a series of bands after agarose gel electrophoresis. The smallest band, the monomer, was estimated to have a length of 240b.p. and the others are multiples of this amount, dimers trimers etc. In addition, there was a second, minor, series of bands which had lengths intermediate between those of the major bands; 120b.p., the half-mer; 360b.p., the one-and-a-half-mer, and so on.

To test whether all satellite sequences are found in each class of restriction fragment, the monomer, the one-and-a-half-mer, the dimer and the trimer were purified from a 1.5% agarose gel and transcribed using α^{32}P-GTP as label. RNAase T_1 fingerprints were made of each of these cRNA classes and the results of the

PLATE VIIa

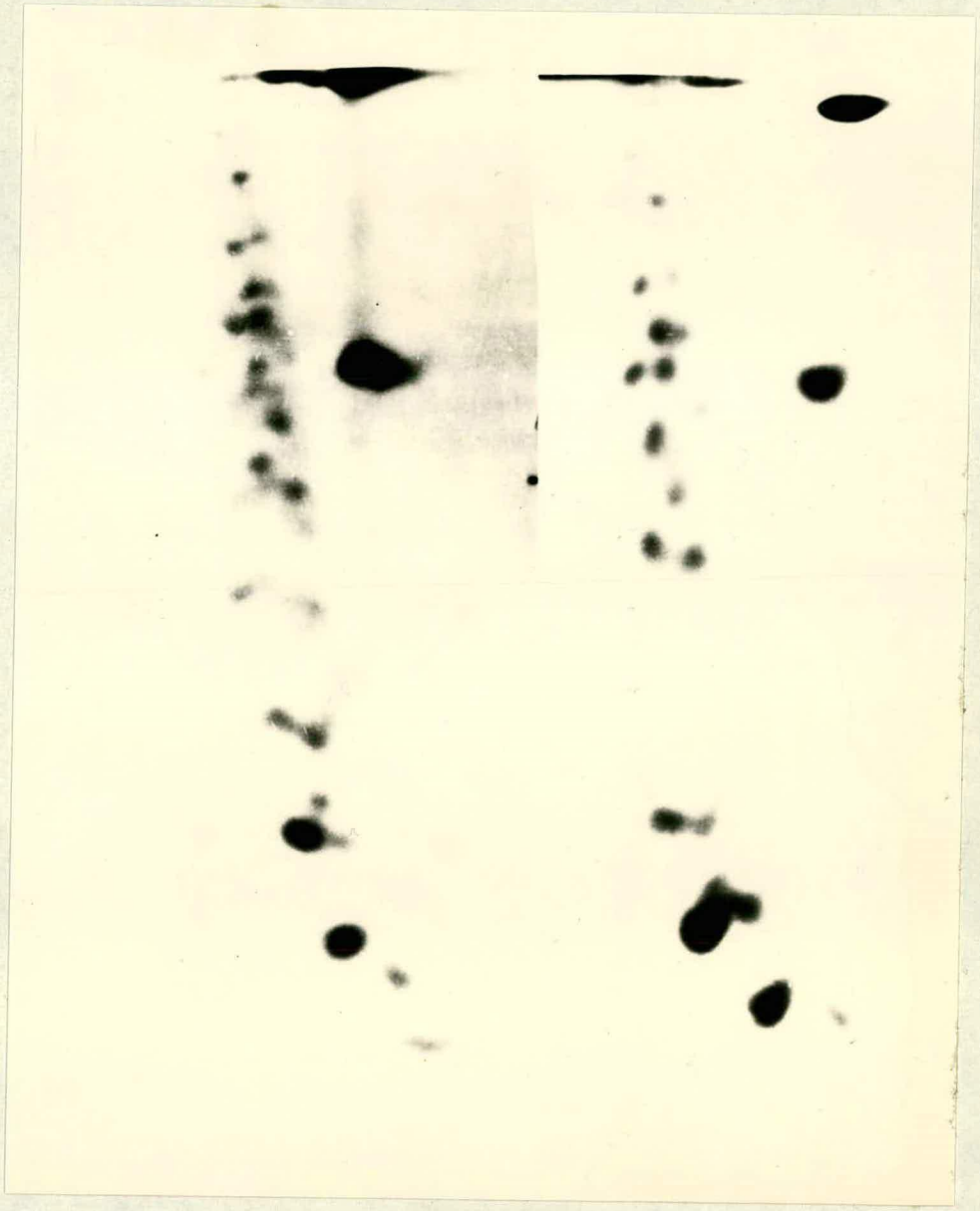
RNAase T_1 Fingerprint of ^{32}P-GTP labelled cRNA transcribed from native satellite DNA.

PLATE VIIb

RNAase T_1 Fingerprints of ^{32}P-GTP labelled cRNA transcribed from:- (a) satellite monomer

(b) satellite trimer

(a) (b)



monomer and trimer are shown in Plate VII and compared with that of a cRNA transcript produced from the same sample of native, undigested satellite and processed at the same time. No significant differences emerge between any of the cRNA classes. All the oligonucleotides which are present in the native cRNA fingerprint are also present, in the same relative amounts, in the fingerprints made from monomer and trimer cRNAs. (The transcripts of the dimer and one-and-a-half-mer cRNAs showed a similar pattern, not shown, but the fingerprints were very faint, owing to lack of material). Hence all classes of restriction fragment contain all the satellite sequences and are fully representative of satellite DNA, and do not represent sequence-distinct subclasses of the DNA.

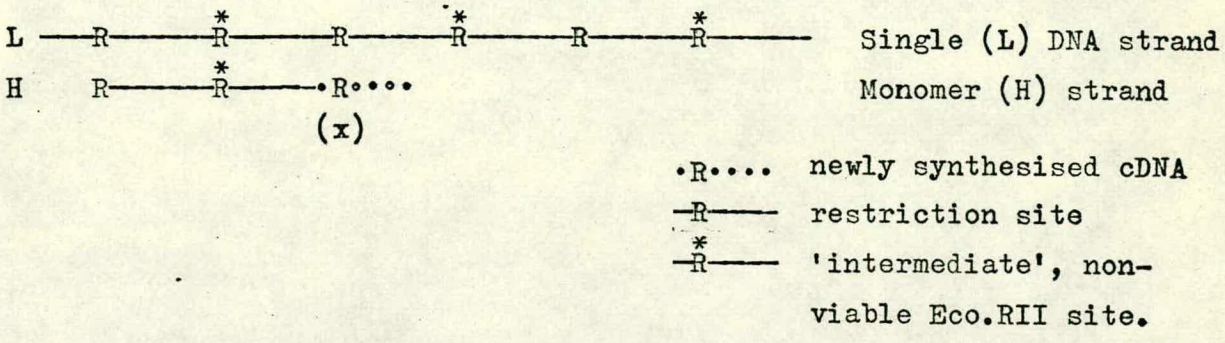
Sequence analysis of cDNA

In view of the results obtained above it was decided to use a total Eco.RII restriction digest of the satellite DNA in the hybridization, rather than a single, purified class of restriction fragment, such as the monomer.

When such a digest is dissociated and allowed to hybridize to 10-fold excess of single strand DNA a number of different reassociation products will be obtained. However, the major product will consist of satellite monomers, (which comprise 70% of the satellite, Southern, 1975), and, to a lesser extent, dimers and trimers etc. hybridized to the complementary single strand of unrestricted DNA. The hybridization may either occur in the 'flush' or in the 'staggered' register, (Southern, 1975). However, only restriction fragments which have reassociated in the 'flush' register will be restricted after extension by *E. coli* DNA polymerase as only such fragments will produce a viable Eco.RII

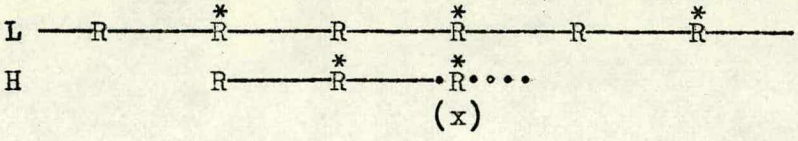
restriction site, see diagram:-

Restriction fragments may reassociate in the 'flush' register:-



in which case, a proper restriction site will be formed after extension by DNA polymerase, at (x)

Or restriction fragments may reassociate in the 'staggered' register:-



in which case, the site produced at (x) after extension by DNA polymerase will not be a true Eco.RII site and the newly synthesised cDNA will not be restricted and hence will not enter the gel.

Consequently, the sequence obtained by this method will only be from one half of the monomer; that half nearest the true restriction site, and will not reflect the whole of the satellite sequence.

In addition to these major products, the restriction fragments will self-reassociate, although only about 10% of the restricted DNA will not be hybridized to the single strand. Some of the fragments will reassociate in the 'flush' register and may be of equal length, in which case they will not be affected by the polymerase. Some may not be of equal length and may give a situation which is formally

equivalent to that produced above: a short fragment may reassociate with a longer fragment in the 'flush' register. Some of these products will consist of a short (L) strand hybridized to a long (H) strand, which is the opposite situation to that produced above and will produce ^{32}P-labelled oligonucleotides from the complementary DNA strand after restriction, and would form bands which would confuse the analysis. However, only a very small proportion of the DNA will form such hybrids and will not provide a major source of error of the sequencing.

Finally, some of the self-reassociated restriction fragments will be in the 'staggered' register and, as such, will not be restricted after polymerase treatment and Eco.RII digestion.

The results of two experiments in which restriction fragments were hybridized to a 10-fold excess of the light strand of the satellite DNA are shown, (Plates VIII and IX), and the sequence obtained is shown in Fig. 4. The sequences obtained from the two experiments are in good general agreement and the most probable sequence of the satellite is shown below in Fig. 4. In both cases the minus system worked for all the nucleotides, whereas the plus system only worked in one case and then only for the +C channel, (Plate VIII). In other experiments, (not shown), the plus system worked over a short stretch of the satellite and confirmed the general analysis, presented below.

A number of observations can readily be made. Firstly, divergence does not seem to greatly affect the separation of the bands. Discrete bands are still visible in the region of over 100 bases and a sequence can be obtained from the radioautograph of the gel. Divergence may lead to a slight increase in band width;

PLATE VIII

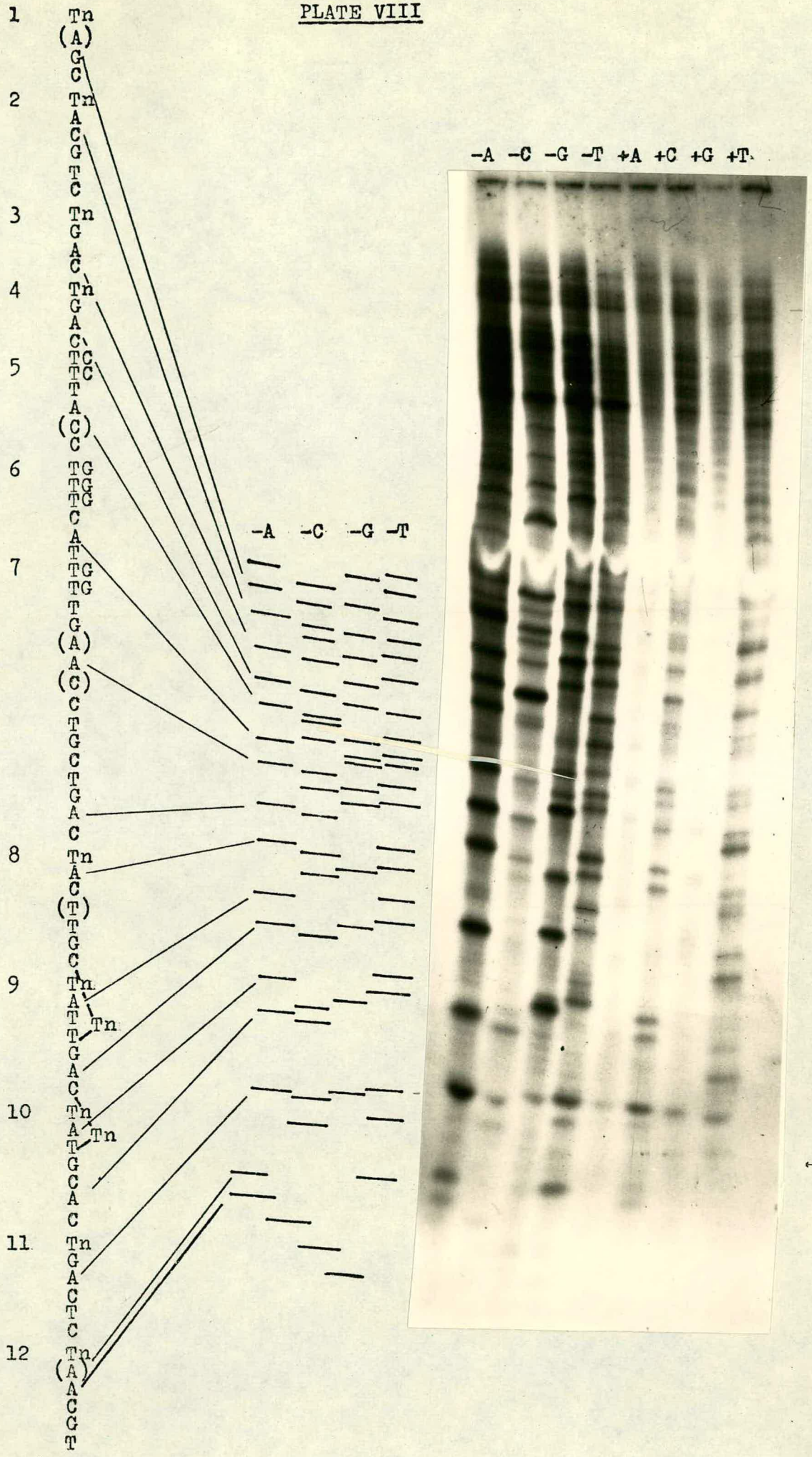
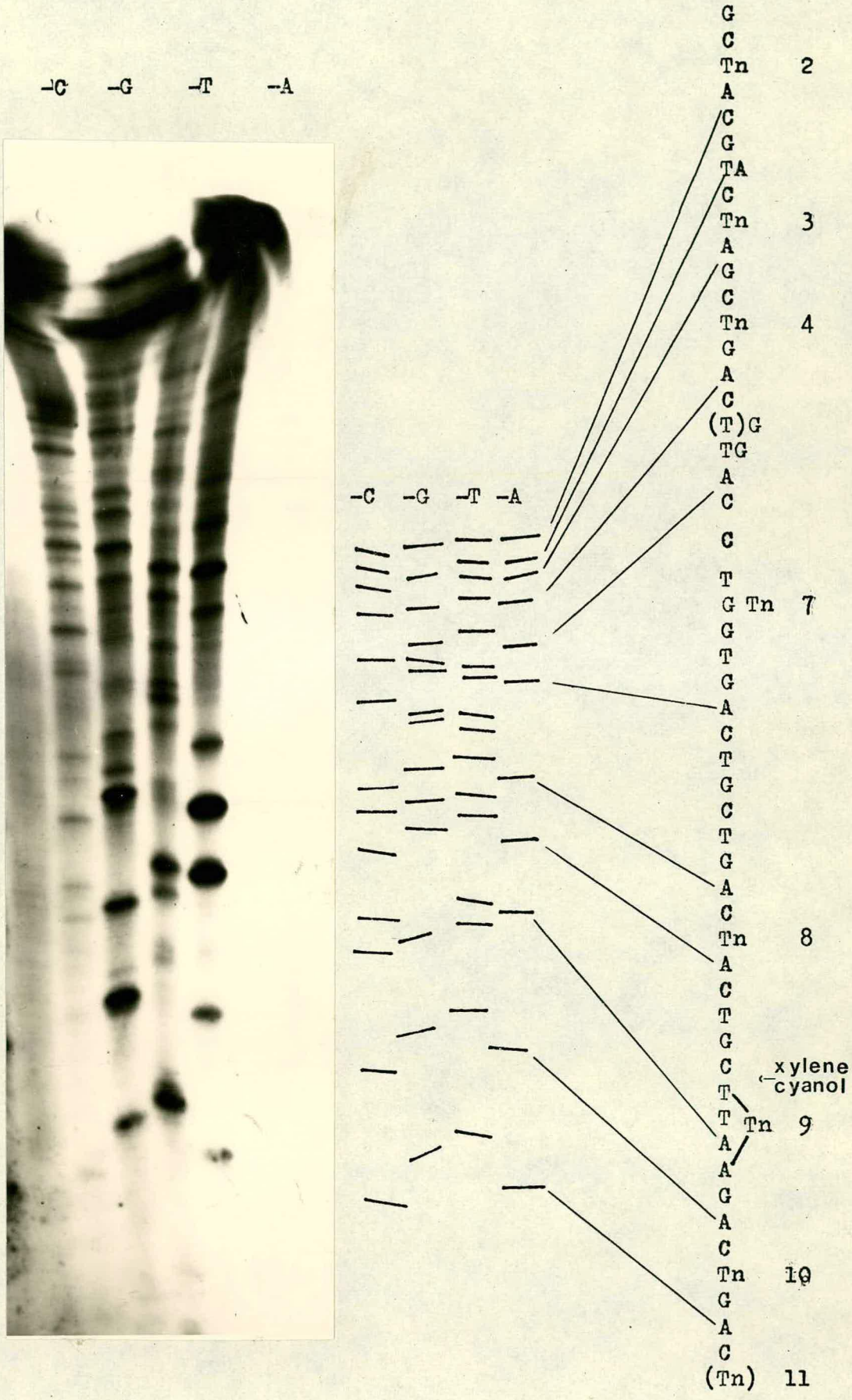


PLATE IX


-C -G -T -A
-C -G -T -A
G
C
Tn 2
A
C
G
TA
C
Tn 3
A
G
C
Tn 4
G
A
C
(T)G
TG
A
C
C
T
G Tn 7
G
T
G
A
C
T
G
C
T
G
A
C
Tn 8
A
C
T
G
C
T
T Tn 9
A
A
G
A
C
Tn 10
G
A
C
(Tn) 11
xylene
cyanol

however, it does appear that this technique can be applied to a diverged sequence such as mouse satellite.

The shortest polynucleotide which is visible in the gel is about 40-50 bases, as judged by the position of the xylene cyanol marker dye, and the longest polynucleotide is over 100 bases. A stretch of about 100 bases, starting 40-50 bases from the RII site, was sequenced. It was not possible to analyse any region closer to the RII site, owing both to the absence of a complete pattern of bands and to the presence of a salt boundary which caused some stacking of bases near the base of the gel. The clear, 'V'-shaped notches seen near the top of each of the channels in Plate VIII are due to the high salt concentration, (50mM NaCl), included in the restriction buffer to inhibit exonuclease activity. Subsequently it was found that omission of the salt from the buffer did not affect the restriction activity of the enzyme or lead to the production of exonuclease activity, (Plate IX).

The short-range repetitious nature of the satellite can clearly be seen in the repeating pattern of a few bands followed by the absence of bands, which mark the T-rich lengths, followed again by the same, or very similar, pattern of a few bands. This is especially evident in the lower half of the gel, and is represented diagramatically below:-

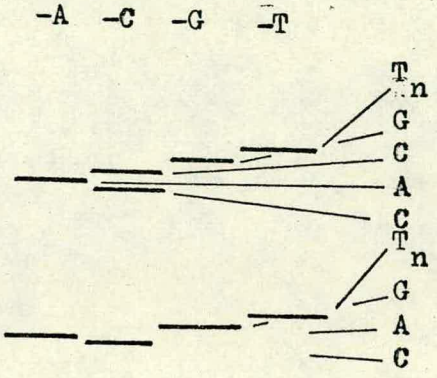
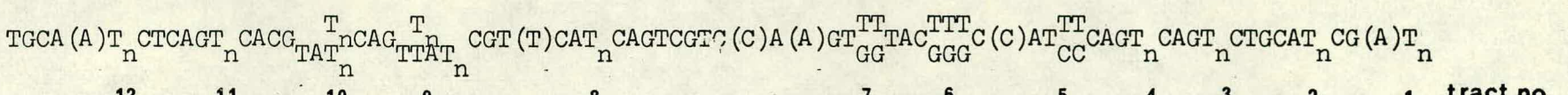
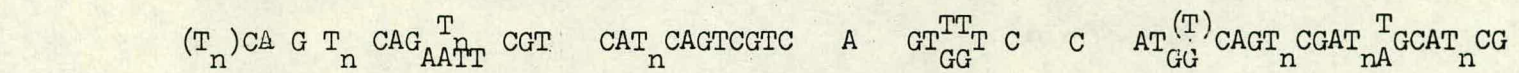


Fig. 4

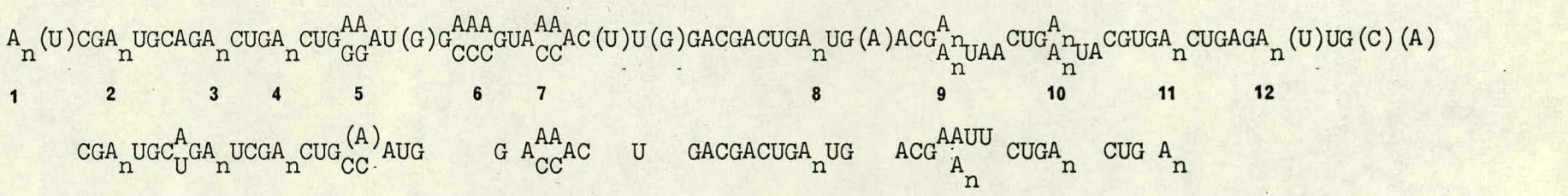
Sequence from Experiment 1



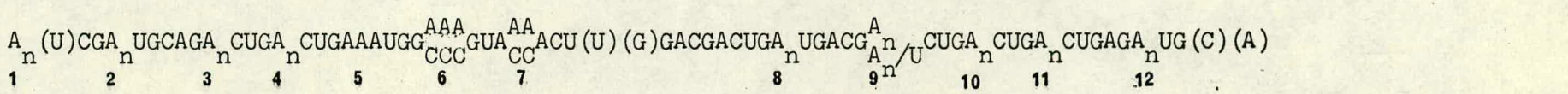
Sequence from Experiment 2



Transposed Sequences:-



Most Probable Sequence:-



Several A-rich tracts can be seen in the analysis shown in Fig. 4 which, although derived from the heavy, T-rich strand has been transposed to that of the complementary, A-rich cRNA strand to facilitate comparison with the results of the cRNA analysis.

Following a series of at least four A-rich tracts, (nos. 1-4, Fig. 4), which follow one another fairly closely, is a less well defined stretch of sequence, (from tract no. 5 to no. 7), which appears to have three runs of A residues, but which may also contain short runs of C residues. Then follows a stretch of about 12 bases which is not related in sequence to the A-rich tracts, but which is rich in G and C bases. However, the precise sequence of this region cannot be given, owing to the presence of bands apparently in the same position, but in different channels. The most probable sequence, judged by the positions of the most intense bands, is shown. Finally, there is a second series of A-rich tracts, nos. 8-12, which may contain minor sequences, especially tracts nos. 9 and, possibly, 10, which may also contain C residues.

The A-rich tracts are mainly of the form:- A_nCUG, rather than A_nUG which was expected on the basis of the cRNA analysis. At least three A_nCUG sequences follow each other with no apparent intermediate sequence, nos. 9-11, and there is also a pair of adjacent A_nCUG tracts, nos. 3-4. However, others are separated by a short sequence: nos. 2-3 are separated by a -CA- pair; nos. 8-9 by a -AC- pair and nos. 11-12 by an -AG- pair. The length of the runs of A residues is 4 or possibly 5, in most cases. Using this technique it is not possible to assign an exact length to runs of the same base, (Sanger and Coulson, 1975).

The origin of the minor bands, seen in the central region from tract no. 6 to no. 9, is discussed later; they may result from

sequence divergence, either base changes in a basic sequence or, more likely, from subclasses of the satellite produced by multiplication of a diverged section of the satellite. Where the sequence is not clear possible alternatives are given. Some of the minor bands may be artifacts, such as those seen by other workers, (Sanger and Coulson, 1975; Air et al. 1975).

The results of the 'plus-and-minus' method confirm that the mouse satellite is composed of related A-rich tracts with possibly short sections of non A-rich sequences, or more highly diverged sequences, occurring at intervals between runs of the A-rich tracts. The A-rich tracts are all of the same general form:- GA_nCUG and could have arisen from the same ancestral sequence.

The sequence presented above could be confirmed by eluting the bands from the gel and performing further analyses such as pyrimidine tract analysis or partial exonuclease degredation. It might also prove possible to use this technique in conjunction with ribosubstitution and carry out two-dimensional fingerprinting on a selection of the eluted, ribosubstituted bands. The data could also be extended by hybridizing the restriction fragments to the complementary DNA strand, which, in this case, would be the heavy strand of the satellite.

Correlation with cRNA analysis.

It has been shown from the cRNA analysis that the satellite has four major A-rich sequences, (spots nos. 21-24 in the $cRNA_H$; A_5UG, A_4CUG, A_4UG and A_3UG), each of which occur, on average, once every 120-150 bases. In addition to these four, there are a number of other sequences related to them by possessing A_nU sequences, where n is at least three. Such sequences account for at least 50%

of the total nucleotides, as measured by the proportion of the radioactivity present in such sequences in the RNAase T_1 fingerprints of the ^{14}C-GTP labelled $cRNA_H$. The estimate of 50% is a minimum, partly for the reasons given in the section on 'Quantitative analysis of the cRNA' and partly because one base change, especially the introduction of a G residue, in an A_nU tract could alter its characteristic sequence pattern such that it may not be recognised as being of the A_nU form.

There is evidence from the cRNA analysis for additional sequences, not related to the A-rich tracts discussed above. Not all the nucleotides can be accounted for by such tracts and the C content of the $cRNA_H$, (14%), and the proportion of monopyrimidines, (about 8% of bases in the DNA are found as monopyrimidines, see Table I) found in the heavy strand DNA seem too high to be accounted for solely by divergence. Two lines of evidence suggested that the additional sequences might occur between the A-rich tracts. Firstly, there are few oligonucleotides which contain two long runs of As, and such repetitive sequences would have been expected to have arisen by a change in the G residue if there was only one G base between them. Secondly, the complementary U-rich spots of the $cRNA_L$ contain 5,6 or 7 U residues but would contain many more if there was no G residue between them; that is, if there was no corresponding C residue between A-rich tracts in the $cRNA_H$.

The length and nature of the intervening sequences was unknown, except that they must contain C and G residues in order to account for the C and G content of the $cRNA_H$, and also of the $cRNA_L$, and may be related to the Eco.RII restriction site sequence: $\frac{CCAGG}{GGTCC}$, (Boyer et al. 1973; Bigger et al. 1973), which occurs once every

240 b.p., (Southern, 1975).

The results of the 'plus-and-minus' method confirm that the satellite DNA contains related A-rich sequences which may have arisen from the same ancestral sequence. However, there are two major contrasting points between the cRNA analysis and this one.

Firstly, the major sequence family found by DNA sequencing is A_nCUG, where n is 4 or 5, whereas in the cRNA analysis, A_nCUG is only one of four related sequences which occur with equal frequency in the $cRNA_H$ fingerprints.

Secondly, these A_nCUG tracts seem to lie adjacent to, or close to one another and often occur in long blocks, not evenly separated by intervening bases. The evidence from the fingerprinting suggested that they may be separated by sequences possibly as long as the A-rich tracts themselves.

Although these points of contrast cannot be resolved absolutely on the available evidence, a number of considerations emerge. The length of the satellite observed by the 'plus-and-minus' method is approximately 100 bases, out of a possible total of 240. Southern, (1975) has shown that the 240 b.p. fragment produced by Eco.RII digestion, (the monomer), is composed of similar halves which, he suggests, may have arisen by a multiplication/duplication event. Furthermore, if the monomer is denatured and allowed to renature it may do so in either the 'flush' or the 'staggered' register, (see previous section). However, only fragments hybridized in the 'flush' register would produce sequenceable products hence the final sequence obtained is from only one half of the 240 b.p.-monomer and consequently constitutes less than half of the total sequence. The other half of the monomer may contain small but significant differences

in sequence, for example, it may contain a higher proportion of A_nUG over A_nCUG tracts. The test of this possibility would be to repeat the experiment using the complementary, (the heavy), DNA strand.

The fidelity and accuracy of the 'plus-and-minus' method has not yet been fully established and it may be that the polymerase makes occassional or recurring mistakes. Kornberg et al. (1964) have shown that DNA polymerase can copy by a slippage mechanism in which the template and the extended primer can slip past one-another and so re-expose the same stretch of template for copying. If such a process occurred in this system some bases, or even short sequences, may be duplicated. Chamberlin and Berg, (1962) also showed that RNA polymerase can transcribe by the same mechanism, in which case the RNA sequencing data could be prone to the same sort of errors. However, the generally accepted fidelity of transcription and the correspondence between the cRNA and the pyrimidine tract data make this a less likely possibility.

Other errors in this analysis may arise. Thus the gel mobility of polynucleotides depends on both length and base composition, so that different sequences of different lengths could co-migrate, instead of being separated on the basis of size, or could be separated by too great a distance: the effect of sequence divergence has not been rigorously investigated. Southern, (1975) has calculated the divergence in the restriction register of the satellite to be 3% and this may affect the separation of the synthesized stretches as these may be up to 150 bases in length and may, therefore differ by up to 5 bases. Although homogeneous polynucleotides which differ only in length give sharp bands on the polyacrylamide gels used, it is possible that the introduction of up to 5 base changes may affect

their separation and lead to errors in the analysis. Such errors would be especially serious if the divergence included additions and deletions as well as base substitutions. The data of Southern, (1975) suggested that the monomer could have a maximum of 10% length heterogeneity, as measured by the band-width on an agarose gel, however, the average length heterogeneity is not known.

In the analysis of the satellite by the 'plus-and-minus' method the effect of heterogeneity may well account for the presence of the minor bands, especially in tracts nos. 9 and 10 which may contain U residues as well as A residues. The $cRNA_H$ analysis also shows heterogeneity in the form of minor spots which contain -UAU- triplets, (spots nos. 5,6,10,14,16 and 18) and some of these spots may be due to the introduction of U residues into the A-rich tracts. This stretch of the sequence was derived from the lower half of the gel where the separation of oligonucleotides is greatest and, consequently, where detail is most clearly visible. Similar areas of divergence may, therefore, be present in the first group of A-rich tracts but they may not be visible in the radioautograph, (tracts nos. 1-4).

Some of the minor bands seen in the 'plus-and-minus' analysis may be produced from small amounts of alternative products of the hybridization reaction, for example from short restriction fragments hybridizing to longer, heavy strand fragments and so producing a cDNA copy of the heavy DNA strand, instead of the light strand, (see previous section). Similarily, some may come from the other half of the same strand of the monomer as a process such as unequal crossing-over could scramble some of the left and right-hand halves of the monomer.

Although the A-rich tracts lie closer to one another than

predicted by the cRNA analysis, they are not immediately adjacent in every case; tracts 2-3, 8-9 and 11-12 are separated by short sequences, (see also previous section) which may contain C residues; between nos. 2-3 and 8-9. Where there are no linking sequences, at least one of the tracts contain a C residue which would produce the necessary G residue in the complementary DNA strand and so account for the absence of very long U-rich oligonucleotides in the RNAase T_1 digests of the $cRNA_L$.

The $cRNA_L$ analysis shows several U-rich spots which end in sequences such as -UCAG, -UCCAG and -UCCACG which correspond to:- CUGA-, CUGGA-, and CGAGGA-, all of which correspond to sequences which link the runs of A in the $cRNA_H$. For example, spots nos. 5, 11, and 15 end in -UCAG, nos. 10, 19 and 20 end in -UCCAG and no. 9 ends in -UCCACG. Hence the results from the $cRNA_L$ analysis are compatible with the proposed sequence deduced from the cDNA data.

The cRNA analysis has revealed several groups of sequences which have different repeat lengths, suggesting there are periodicities within the satellite. However, the 'plus-and-minus' method examines the sequence at only one level, the 240 b.p. level, as there is a restriction site every 240 b.p. Consequently, the minor sequences, which can be seen in the cRNA analysis, although they may produce visible bands in the gels of the cDNA analysis cannot be analysed by this method which can only be used to determine the predominant sequence pattern.

CHAPTER FOUR

Arrangement and Evolution of the Sequence

Southern, (1970), has discussed the evolution of the Guinea pig satellite I. He showed it was based on a hexanucleotide repeat which, he suggested, had diverged and subsequently evolved through a series of discrete multiplication and divergence events. The results described above support the view that the mouse satellite, like several others, including the Guinea pig satellite I, is composed of a diverged, short, repeating sequence.

The cRNA analysis reveals four closely related sequences, based on an original A-rich tract such as (G)A_5UG(A) or (G)A_4CUG(A), and each of these four sequences occur, on average, once every 120-150 b.p. in the satellite. The DNA sequencing data suggests that the original repeat may have had a sequence such as ...GA_4CUGAC... or ...GA_5UGAC... and may have been about 10 b.p. in length which is in keeping with the estimate of Southern, (1970), based on pyrimidine tract analysis.

The quantitative data obtained from the cRNA analysis show that sequences related to the A-rich tracts account for 50% of the total nucleotides, hence a maximum estimate for the length of the repeat would be twice the length of an A-rich tract, about 18 b.p. (Biro et al. 1975). However, this estimate makes a number of assumptions about the organization of the sequence. Firstly, it assumes that the spots observed after RNAase T_1 digestion are representative of the whole satellite, which may not necessarily be the case; it is possible that the RNA polymerase initiates preferentially at certain sites in the DNA, (see section on RNA sequencing). Secondly, it assumes that additional nucleotides are

uniformly interspersed between the A-rich tracts. Although there is evidence from the DNA sequencing data to suggest that some additional nucleotides, including at least one C residue, do lie between some of the A-rich tracts, the DNA analysis suggests that these additional sequences may also occur in longer runs of about 12-15 bases. A scheme for the evolution of the mouse satellite can now be derived from the sequencing data presented above and from the restriction data of Southern, (1975).

The simplest scheme to fit the structure would be that the original 12 b.p. ancestral sequence was multiplied and subsequently underwent divergence. At a later multiplication stage the four A-rich sequences seen in the $cRNA_H$ analysis were formed. A minimum estimate of the extent of the divergence which occurred between the two stages can be calculated. The least number of base changes needed to derive all four sequences, (GA_5UGA, GA_4CUGA, GA_4UGA and GA_3UG; 32 bases in all), from a common ancestor is four, if GA_5UGA is assumed to have been the original sequence. Hence, at least 12% divergence must have occurred, (4 base changes in 32), to give rise to the four major sequences.

The length of the sequence at the second round of multiplication would have been about 50 b.p.. But, the present day frequency of each of these four sequences is once per 120 b.p., not once per 50 b.p. Hence, at least 10% further divergence must have occurred since that second round of multiplication in order to halve their frequency, using the equation: $x = (100-p)^n$, where x is the proportion of oligonucleotides of length n b.p. which remain unchanged after p percentage divergence to calculate p.

From the two estimates, the total amount of divergence in the ancestral sequence must have been over 20%. Rice, (1972), has estimated that rodent repeated sequence DNA evolves at a rate of

1% of base changes per 10^6 years. Therefore, the original multiplication event probably occurred about 20×10^6 years ago, and the second event took place 10×10^6 years ago.

Sutton and McCallum, (1972) showed that cross-reassociated M.mus and M.caroli DNA satellites had a 20°C reduction of Tm which is equivalent to between 16 and 20% mismatching, depending on the conversion factor used, (see Introduction - reassociation), and this is close to the 20% estimated from sequencing data. Taken together, these results suggest that M.mus and M.caroli satellites diverged at the same time as, or very shortly after, the first multiplication event - 20×10^6 years ago, and consequently share the same ancestral sequence. Analysis of M.caroli satellite DNA may shed further light on the nature of the sequence. Sutton and McCallum, 1972), showed that the rates of cross-reassociation of M.mus and M.caroli satellites were very high, considering the high degree of mismatching involved and they estimated the sequence complexity of the DNA recognized in the cross-reassociation reactions to be only 8-20 b.p. which is in the same range as the estimate from sequencing. In homologous reassociation reactions, performed under identical conditions, the length of the DNA sequence which is being recognized would be the 120 b.p. unit which Sutton and McCallum, (1971), quote as the complexity of M.mus satellite DNA. In heterologous reassociation reactions, however, this register would not be recognized owing to the high degree of divergence between the species.

Southern, (1975), has proposed third and fourth stages of multiplication. Evidence for a third stage, involving a sequence about 120 b.p. long, which would comprise two of the 50 b.p. stretches, comes from the analysis of restriction fragments. From the present analysis, it is possible that this may have included

some atypical DNA, possibly 10-20 b.p. in length, which would account for the short stretch of non A-rich nucleotides seen by the 'plus-and-minus' method. If introduced at this stage, such a foreign sequence, not derived from the original starter sequence, would occur every 120 b.p.; however, it could have been introduced at the next stage, the fourth round of multiplication, in which case it would occur every 240 b.p.

Alternatively, these atypical sequences could be accounted for by an unusually high rate of divergence in some stretches of the satellite and have been spread by a mechanism such as unequal crossing-over, (see subsequent section).

Hutton and Wetmur, (1973), have also proposed a model for mouse satellite DNA which included regularily spaced regions of non-satellite DNA lying between short satellite sequences which are less than 36 ± 9 b.p. long. However, the results of the 'plus-and-minus' method clearly show that the characteristic, A-rich, satellite sequences must be longer than the 45 b.p. maximum allowed by Hutton and Wetmur as five or more A-rich tracts are seen to lie adjacent to one another.

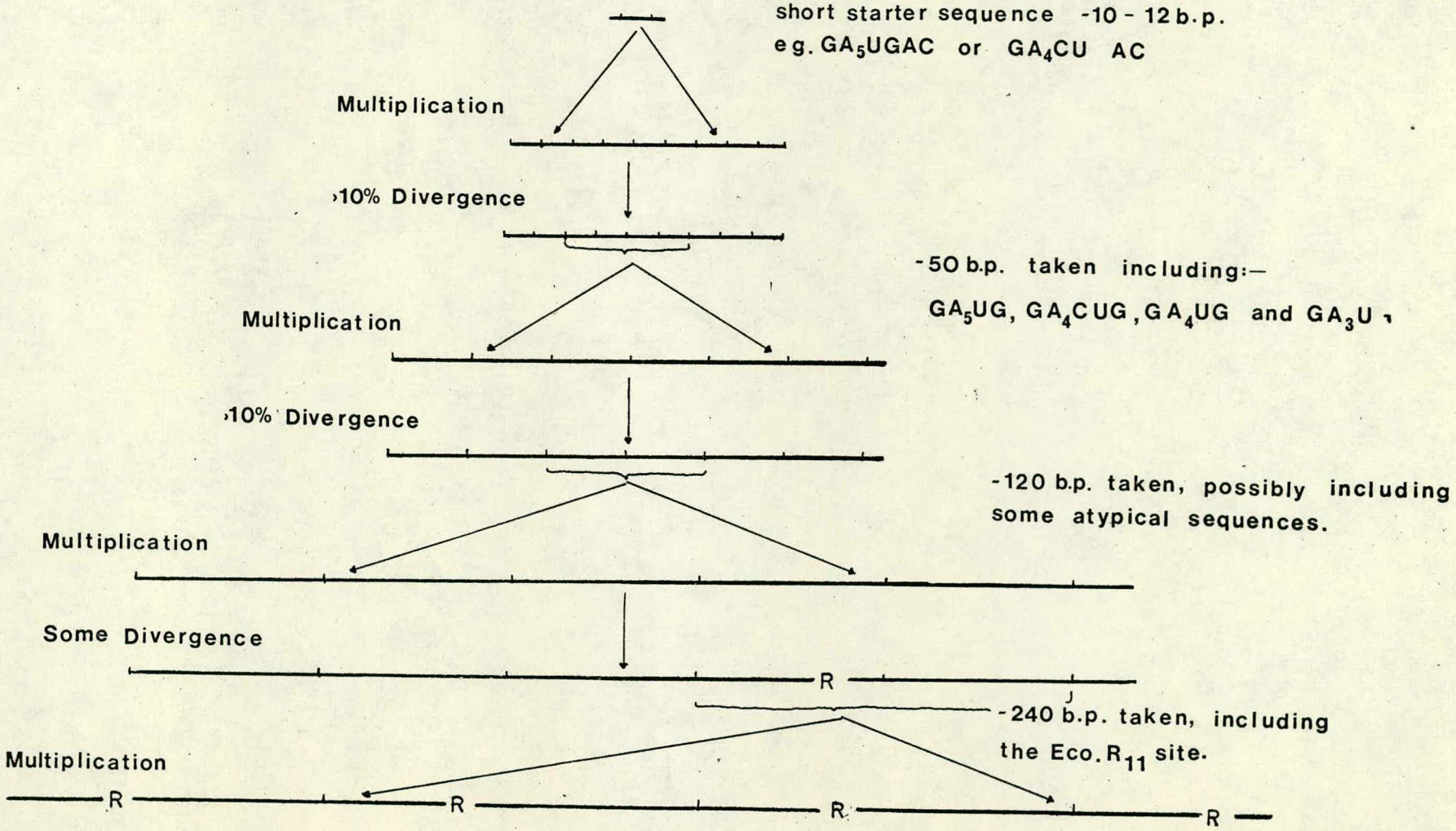
The fourth round of multiplication was proposed to account for the restriction register of 240 b.p. and must have taken place not long after the third round as Southern, (1975), has shown that reassociated restriction fragments have about the same reduction of Tm, ($5^{\circ}C$), as reassociated, unrestricted satellite. Hence the reassociation register, (120 b.p.), and the restriction register, (240 b.p.), have about the same degree of divergence which is about 3%. Therefore, the third round of multiplication probably took place about $3\text{-}5 \times 10^6$ years ago and the fourth round not long after, probably about 3×10^6 years ago.

Although coding DNA evolves less rapidly than non-coding DNA, (McCarthy and Farquar, 1972; Rosbash et al. 1975), owing to greater selection pressure on coding sequences, the bulk of single copy DNA evolves at the same rate as repeated sequence DNA, (see Southern, 1974). Rice, (1971 and 1972) and Rice and Straus, (1973), have shown that cross-reassociated single-copy DNA from M. mus and M. caroli also have a 5°C reduction of Tm and this has led Southern to suggest that the third round of multiplication took place at the same time as speciation, that is when the two Mus species separated, while Sutton and McCallum had already shown that the ancestral satellite sequences, which are common to the two species, (Sutton and McCallum, 1972; Southern, 1975), were formed well before speciation.

The schème for the evolution of the mouse satellite is shown diagrammatically in Fig 5, and is based on that shown in Southern, (1975).

Fig. 5

Scheme for the Evolution of Mouse Satellite DNA



Mechanisms for the Formation of Tandemly Repeated Sequences

The multiplication model, described above, makes no assumption about the molecular mechanisms involved in the generation of multiple copies of defined sequences. The presence of more than one periodicity within the satellite rules out the possibility that the satellite evolved by a single, sudden multiplication event which gave rise to all the satellite at one stage, and possibily in one individual, such as originally proposed by Britten and Kohne, (1969).

Most of the discussion on the evolution of repeated sequences has centered on those genes which are known to be present in the germ line in multiple copies, (multigene families, Hood et al. 1975; or tandons, Tartof, 1975). Tartof defines tandons as segments of DNA which contain a collection of similar or identical repeating units. It would include the antibody 'V' genes, (Cohn, 1974), sea urchin histone genes, (Kedes and Birnstiel, 1971), Drosophila 5s genes, (Wimber and Steffensen, 1970), Xenopus 5s genes, (Brown and Weber, 1968), and ribosomal and tRNA genes in ampbibia, (Birnstiel et al. 1971 and 1972), and also in Drosophila, (Ritossa et al. 1966a). The definition of tandons also includes the satellite DNAs. Some genes also undergo somatic amplification during the developement of the organism, giving rise to the phenomenon of ribosomal gene amplification in Drosophila, (Lima de Faria et al. 1969), and in Xenopus, (Brown and Dawid, 1968), and also to the phenomenon of chromosomal polyteneization in the Diptera, (Gall et al. 1971).

Consequently, assuming such multiple genes are descended from a common ancestor, mechanisms must exist for the multiplication of DNA, both during somatic developement and in the germ line. Two

general theories which have been advanced to explain these processes will be considered here.

The first was put forward by Gilbert and Dressler, (1968) to to account for bacterial and phage chromosome replication and has been termed 'the rolling circle model'. The DNA which is to be replicated forms a nicked, closed circle. Synthesis commences at the 3' end of the nicked strand which thus becomes the primer and uses the complementary, closed strand as template. The newly replicated strand is reeled off the template in a continuous process and may, at least in theory, contain an unlimited number of copies of the template. The complementary DNA strand is then made on the newly replicated strand which is used as the template.

Hourcade et al. (1973a and b) and Bird et al. (1973) have detected such circles of DNA by electronmicroscopy in Xenopus oocytes and have concluded that this mechanism is responsible for rDNA amplification in Xenopus. However, such DNA is not integrated into the chromosome and does not become part of the organism's genome.

Ritossa, (1972), has suggested that Drosophila rDNA genes may be multiplied by a rolling circle process. Although rDNA genes are present in multiple copies in Drosophila, mutants are found which are partially deficient for these genes, the bobbed, (bb), mutation, (Ritossa et al. 1966b). An X chromosome bearing a bobbed mutation can revert to wild type if it is maintained in the heterozygous state for a few generations. This process occurs by the accumulation of additional genes and has been termed gene magnification, (Henderson and Ritossa, 1970).

Support for this view comes from recent experiments of Locker, (1976), who produced evidence to suggest that the wild type

revertants of the bobbed mutation are genetically unstable and may loose their extra DNA after one or two generations. Hence Locker proposes that the magnified DNA is not covalently linked to the chromosomal DNA and, therefore, an extra-chromosomal mechanism, such as that proposed by Ritossa, must be responsible for the magnification phenomenon.

Although there is no direct evidence about satellite DNA, Botchan, (1974), favours this model to account for the evolution of the bovine satellites.

The second mechanism proposed to account for repeated sequence DNA formation is one of unequal mitotic sister chromatid exchange. Tartof, (1973), has shown that when a mutant bobbed Y chromosome in a heterozygous male Drosophila reverts to wild type, (magnification), there is a corresponding loss of rDNA genes in the wild type X chromosome, (reduction). Tartof also used an X-ring chromosome, in which the frequency of crossing-over is greatly reduced, to show that the frequency of the magnification/reduction process is similarily reduced. Hence he proposes that unequal mitotic crossing-over is responsible for rDNA magnification in Drosophila and may also be responsible for maintaining homogeneity among other repeated gene families and satellites.

The Tartof model is a general one and Smith, (1973), and also Black and Gibson, (1974), have performed computer simulation studies to show that unequal crossing-over between sister chromatids can account for both homogeneity of repeated sequences and the spread of mutations among adjacent sequences.

Smith showed that if a stretch of 500 copies of the same sequence is permitted to undergo unequal sister-chromatid exchange

such that the length variation of the products is within 10% of the original 500 copies, then, after 10^4 such cross-over events, all the resulting copies will be descended from only one copy of the original 500. The degree of homogeneity between the final sequences will depend on the relative rates of crossing-over and mutation, and also on selective pressures, if present. Smith argues that it could account for the evolution of rDNA genes and spacers, immunoglobulin 'V' genes and also satellites. Smith estimates that, assuming a mutation rate of 5×10^{-9} mutations/b.p./generation and 1 cross-over per generation, then 10^4 cross-overs will be required to fix 500 rDNA genes to homogeneity. If the mutation rate is low compared with the fixation rate, then the sequence will be very homogeneous, (Cf. Drosophila satellites). However, if the mutation rate, or the number of copies of the sequence, is large, then the resulting sequence will be less homogeneous and mutations introduced into a single copy of the sequence will spread to adjacent copies, giving rise to sub-families.

Support for Smith's model comes from studies on Xenopus 5s and ribosomal gene spacers. Wellauer and Dawid, (1974), have shown by heteroduplex mapping and by restriction analysis, (Wellauer et al. 1974), that the non-transcribed ribosomal spacers, (Brown and Weber, 1968), of amplified Xenopus laevis DNA are heterogeneous in length. Brown and Wellauer, (unpublished results, quoted in Hood et al. 1975 and in Tartof, 1975) have now cloned restriction fragments of the spacers and have analysed them by both homoduplex and by heteroduplex mapping and have demonstrated the existence of mismatched loops, of constant length, (slightly under 50 b.p.), or of multiples of that amount, in the spacer regions of the reassociated duplexes. It is concluded from this that the spacer contains a series of short

tandemly repeated, homologous units which vary in number from one spacer to the next.

Length heterogeneity has also been found in Drosophila melanogaster rDNA spacers, (Hamkalo et al. 1973), although it may be that the heterogeneity is between different rDNA families and that all spacers within the same family are homogeneous, (Laird and Chooi, quoted in Tartof, 1975).

A similar situation exists in the spacers of 5s DNA. Brownlee et al. (1972 and 1974) have performed sequence analysis on 5s DNA and have demonstrated heterogeneity in the spacers and have suggested that the spacers are composed of a tandemly repeated, heterogeneous sequence which is based on a 15 b.p. unit repeat. Although, as there are families for 5s genes on 15 separate chromosomes, (Pardue et al. 1973), the heterogeneity could have been between families, rather than within them, (Tartof, 1975).

However, evidence for 'scrambled' length heterogeneity of 5s spacers comes from restriction experiments of Brown, (unpublished results). Hind.III cleavage of 5s DNA from individual frogs produces a series of bands which differ from one another in length by about 15 b.p. which is the length of the individual unit repeat of the spacer sequence. Partial digests of the 5s DNA were made and 4-repeat unit fragments were cloned and redigested with Hind.III and these were found to produce heterogeneous length fragments.

Brown interprets these results as supporting Smith's model, as a sudden amplification mechanism, such as by a 'rolling circle' process would produce identical adjacent repeats, which would not be produced by unequal crossing-over.

Cooke, (1975), has suggested that unequal crossing-over, or a

combination of a sudden multiplication event followed by unequal crossing-over, could account for the long-range sequence relationships observed in the satellites of different species of the genus Apodemus, (Allan, 1974; Cooke, 1975).

Southern, (1975), has suggested that unequal crossing-over could account for some of the features of mouse satellite evolution. He proposes that multiple periodicities could have arisen because of changes in the length of the sequence necessary for a cross-over event. As mutations accumulated in short satellite repeats, a longer stretch of DNA would be required to provide sufficient homology for pairing and the switch from a shorter to a longer register would have occurred. Southern argues that unequal crossing-over could account for non-integral spacing of the Eco.RII site and for the non-random distribution of the Hae.III and Hind. sites.

In a subsequent paper, Smith, (1976), has expanded his model by showing that if a random stretch of DNA is allowed to accumulate mutations and undergo unequal crossing-over it will develop sequence periodicities. In his computer simulation model he envisages a stretch of 500 b.p. and allows it to accumulate one mutation per generation and up to 500 attempts at crossing-over. Four base pairs are required to match for a cross-over to succeed and the length of the products is again kept to within 10% of the starting sequence. After 200 cycles, periodicities, usually between 11 and 30 b.p. in length, formed in nearly all the tests. The ends of the sequence, however, remained less homogeneous, owing to the smaller probability of sequence homology, and hence crossing-over, between a short terminal sequence and an internal stretch. This picture is reminiscent of the atypical sequences found among the RNAase T_1 digestion

products of the cRNA.

Long periodicities generated by such a process would, Smith suggests, be unstable and tend to form shorter ones, owing to the formation of points of homology at different places within the longer periodicity. This would result in out-of-register cross-overs within the repeat and thus generate a shorter final periodicity. Smith suggests that this process could account for the formation of the short-range periodicities of mouse satellite. Alternatively, but less likely, the related, but non-homologous, subrepeats so produced could accumulate sufficient mutations to prevent internal, out-of-register crossing-over and this would stabilize the longer register periodicity.

However, a more probable method of generating longer-range periodicities, suggested by Southern, (1975), starts from a homogeneous repeated sequence composed of short periodicities which switches to a longer register, owing to the accumulation of mutations. Smith has tested this by computer simulation and showed that a homogeneous 5 b.p. repeated sequence formed longer, 10 b.p. periodicities which were based on two similar, but not identical, 5 b.p. units. In this, and in other tests, there was a switch in cross-over register from a shorter to a longer unit. These longer periodicities may, of course, form shorter ones, by the mechanism described above and, therefore, Smith suggests, satellites may be in a state of continuous fluctuation in complexity during evolution and be heterogeneous during the transition periods. Such a model could explain both the homogeneity of some satellites, such as those of Drosophila, and also the heterogeneity of others, such as the Guines pig satellite I and the mouse satellite, without the need for any special correction mechanisms.

Once a long-range periodicity has been formed, it will tend to destabilize and form shorter periodicities. However, long-range periodicities are observed in mouse satellite DNA and in bovine satellites, (Botchan, 1974). Smith suggests that such long-range periodicities may be stabilized if the number of matching base-pairs required for a successful cross-over event is increased. Smith finds that longer-range periodicities are formed if 8 b.p. are required for a cross-over than if 4 b.p. are required. However, he was unable to test his model, owing to the large number of calculations which would have been required to generate periodicities of 100-1000 b.p. in a suitable length of DNA. Nonetheless, Smith suggests that unequal crossing-over can account for the formation of long-range periodicities observed in the mouse and calf satellites.

Finally, Smith suggests unequal crossing-over alone can account for sequence homogeneity in the mouse satellite DNA. He assumes a mutation rate of 5×10^{-9} mutations/b.p./year, (Kimura, 1973), which is 0.5% per 10^6 years and which is of the same order as the estimate of Rice, (1972), which was 1% per 10^6 years. To counteract this, 0.1 cross-overs in the satellite per cell per mitotic generation are required to maintain homogeneity amongst the 10^6 copies of the satellite monomer, assuming one organismal and 20 germ-line mitotic generations per year. The total number of cross-overs would then be 20×10^6, which, he concludes, is acceptable.

The unequal cross-over models only take into account crossing-over between sister chromatids, not between chromosomes. Mouse satellite DNA is found on all mouse autosomes and, as exchange events between non-homologous chromosomes are, presumably, rarer than between sister chromatids, the satellite DNAs from different chromosomes may contain sequence differences. Thus some of the

heterogeneity detected by sequencing could be due to 'sub-families' of satellite sequences on different chromosomes and such 'sub-families' could also account for the distribution of the Hae. and Hind. restriction sites.

Tartof, (1975), has described a similar model to the Smith one but Tartof also assumes branch migration may occur after the formation of the cross-over point, in which case, the number of cross-over events required to fix a mutation to homogeneity would be reduced. Tartof estimates that a pentanucleotide repeated 10^5 times would require 10^3 cross-over events, under conditions of branch migration of up to 2000 b.p., whereas the Smith model would predict 2×10^6 cross-overs would be needed.

Although direct experimental evidence for any of these models is lacking, the unequal-crossing over models have been sufficiently well developed to be able to account for all the aspects of mouse satellite DNA evolution, as determined by the sequencing data, presented in this thesis, and by the restriction analysis of Southern, (1975). The multiplication stages, outlined in the previous section and in Fig. 5, could represent changes in the crossing-over register. Such a detailed analysis has not, so far, been performed for the rolling-circle model.

REFERENCES

Adams, J.M., Jeppesen, P.G.N., Sanger, F. and Barrell, B.G. (1969). Nature, 223, 1009-1014.

Air, G.M., Blackburn, E.H., Sanger, F. and Coulson, A.R. (1975). J. Molec. Biol. 96, 703-719.

Allan, J. (1974). Ph.D. Thesis, University of Edinburgh.

Arrighi, F.E., Mandel, M., Bergendahl, J. and Hsu, T.C. (1970). Biochem. Genetics, 4, 367-376.

Bachman, K., Goin, O.B. and Goin, C.J. (1972). In 'Evolution of Genetic Systems', Brookhaven Symposium, (Smith, H.H. Ed.), pp.419-450, Gordon and Breach, New York, London and Paris.

Beattie, W.G. and Skinner, D.M. (1972). Biochim. Biophys. Acta 281, 169-178.

Berg, P., Fancher, H. and Chamberlin, M. (1963). In 'Symposium on Informational Macromolecules', pp.467-483, Academic Press, New York and London.

Bernardi, G. (1971). In 'Methods in Enzymology' (Grossman, L. and Moldave, K. Eds.), vol.XXI, pp.95-139, Academic Press, New York and London.

Bernardi, G., Ehrlich, S.D. and Thiery, J-P., (1974). In 'Methods in Enzymology' (Grossman, L. and Moldave, K. Eds.), vol. XXIX, pp.341-355, Academic Press, New York and London.

Bigger, C.H., Murray, K. and Murray, N.E. (1973). Nature New Biology, 244, 7-10

Bird, A., Rochaix, J. and Bakken, A. (1973). In 'Molecular Cytogenetics' (Hamlako, B.A. and Papaconstantinou, J. Eds.), pp.49-58, Plenum Press, New York.

Birnstiel, M.L., Spiers, J., Purdom, I.F., Jones, K.W. and Loening, U. (1968). Nature, 219, 454-463.

Birnstiel, M.L., Chipchase, M. and Spiers, J. (1971). Prog. Nucleic Acid Res. and Molec. Biol., 11, 351-389.

Birnstiel, M.L., Sells, B.H. and Purdom, I.F. (1972). J. Molec. Biol., 63, 21-39.

Biro, P.A., Carr-Brown, A., Southern, E.M. and Walker, P.M.B. (1975). J. Molec. Biol., 94, 71-86.

Black, J.A. and Gibson, D. (1974). Nature, 250, 327-328.

Blumenfeld, M. (1973). Cold Spring Harbor Symp. Quant. Biol., 38 423-427.

Bolton, E.T., Britten, R.J., Cowie, D.B., Roberts, R.B. Szafranski, P. and Waring, M.W. (1965). Carnegie Inst. Science Year Book, 64, 313-345.

Bonner, T.I., Brenner, D.J., Neufeld, B.R. and Britten, R.J. (1973). J. Molec Biol., 81, 123-135.

Bostock, C.J. (1972). Advances in Cell Biology, (Prescott, D.M. and Goldstein, L. and McConkey, E.M. Eds.), 2, pp.153-223, Appleton-Century-Crofts, New York.

Bostock, C.J. and Prescott, D.M. (1971). Exp. Cell Res. 64, 267-274.

Bostock, C.J., Prescott, D.M. and Hatch, F.T. (1972). Exp. Cell. Res. 74, 487-495.

Botchan, M. (1974). Nature, 251, 288-292.

Botchan, M., McKenna, G. and Sharp, P.A. (1973). Cold Spring Harbor Symp. Quant. Biol. 38, 383-395.

Boyer, H.W., Char, L.T., Dugaiczyk, A., Hedgepeth, J. and Goodman, H.M. (1973). Nature, New Biology, 244, 40-43.

Britten, R.J. (1969). Carnegie Inst. Science Year Book, 67, 332.

Britten, R.J. (1972). In 'Evolution of Genetic Systems', Brookhaven

Symposium, (Smith, H.H. Ed.), pp.80-94, Gordon and Breach, New York, London and Paris.

Britten, R.J. and Davidson, E.H. (1969). Science, 165, 349-357.

Britten, R.J. and Davidson, E.H. (1971). Quart. Rev. Biol. 46, 111-133.

Britten, R.J. and Kohne, D.E. (1967). Carnegie Inst. Science Year Book, 65, 78-106.

Britten, R.J. and Kohne, E.H. (1968). Science, 161, 529-540.

Britten, R.J. and Smith, J. (1970). Carnegie Inst. Science Year Book, 68, 378-386.

Brown, D.D. and Dawid, I.B. (1968). Science, 160, 272-280.

Brown, D.D. and Weber, C.S. (1968). J. Molec. Biol. 34, 661-680.

Brown, D.D., Wensink, P.C. and Jordan, E. (1971). Proc. Natl. Acad. Sci. 68, 3175-3179.

Brown, D.D., Wensink, P.C. and Jordan, E. (1972). J. Molec. Biol. 63, 57-73.

Brownlee, G.G. (1972). In 'Laboratory Techniques in Biochemistry and Molecular Biology', (Work, T.S. and Work, E. Eds.), vol. 3 part I, North-Holland, Amsterdam.

Brownlee, G.G., Cartwright, E.M. and Brown, D.D. (1974). J. Molec. Biol. 81, 703-718.

Brownlee, G.G., Cartwright, E.M., McShane, T. and Williamson, R. (1972). F.E.B.S. Letts. 25, 8-12.

Brownlee, G.G. and Sanger, F. (1967). J. Molec. Biol. 23, 337-353.

Brownlee, G.G. and Sanger, F. (1969). E. J. Biochem. 11, 395-399.

Buongiorno-Nardelli, M. and Amaldi, F. (1970). Nature, 225, 946-948.

Burgess, R.R. (1969). J. Biol. Chem. 244, 6160-6167.

Burton, K. and Petersen, G.B. (1960). Biochem. J. 75, 17-27.

Cech, T.R. Rosenfeld, A. and Hearst, J.E. (1973). J. Molec. Biol. 81, 299-325.

Chamberlin, M. and Berg, P. (1962). Proc. Natl. Acad. Sci. 48, 81.

Cheng, T-Y. and Sueoka, N. (1961). J. Molec. Biol. 4, 161-172.

Chilton, M-D. (1973). Nature, New Biol. 246, 16-17.

Cohn, M. (1974). In 'Progress in Immunology' (Brent, J. and Holborow, J. Eds.) II, 261-284.

Cooke, H.J. (1975). J. Molec. Biol. 94, 87-99.

Corneo, G., Ginelli, E., Soave, C. and Bernardi, G. (1968a). Biochemistry 7, 4373-4379.

Corneo, G., Ginelli, E. and Polli, E. (1968b). J. Molec. Biol. 33. 331-335.

Corneo, G., Ginelli, E. and Polli, E. (1970a). J. Molec. Biol. 48, 319-327.

Corneo, G., Ginelli, E. and Polli, E. (1970b). Biochemistry 9, 1565-1571.

Coudray, Y., Quetier, F. and Guille, E. (1970). Biochim. Biophys. Acta, 217, 259-267.

Darlington, C.D. (1937). In 'Recent Advances in Cytology' 2nd. edition, Blakiston, Philadelphia, Penn.

Davidson, E.H. and Britten, R.J. (1973). Quart. Rev. Biol. 48, 565-613.

Dev, V.G., Miller, D.A., Tantravaki, R., Schreck, R.R., Roderick, T. H., Erlanger, B.F. and Miller, O.J. (1975). Chromosoma, 53, 335-344.

Eckhart, R.A. and Gall, J.G. (1970). Chromosoma, 32, 407-427.

Edström, J.E. and Lambert, B. (1975). Prog. in Biophys. and Molec. Biol. 30, 57-82.

Endow, S.A., Polan, M.L. and Gall, J.G. (1975). J. Molec. Biol. 96, 665-692.

Filipski, J., Thiery, J-P. and Bernardi. G. (1973). J. Molec. Biol. 80, 177-197.

Fisher, W.D., Cline, G.B. and Anderson, N.G. (1964). Analyt. Biochem. 9, 477-492.

Flamm, W.G. (1972). Int. Rev. Cytology, 32, 1-51.

Flamm, W.G., Bernheim, N.J. and Brubaker, P.E. (1971), Exp. Cell Res. 64, 97-104.

Flamm, W.G., Bond, H.E., Burr, H.E. and Bond, S.B. (1966a). Biochim. Biophys. Acta, 123, 652-654.

Flamm, W.G., Bond, H.E., Burr, H.E. and Bond, S.B. (1966b). Biochim. Biophys. Acta, 129, 310-317.

Flamm, W.G., McCallum, M. and Walker, P.M.B. (1967). Proc. Natl. Acad. Sci. 57, 1729-1734.

Flamm, W.G. McCallum, M. and Walker, P.M.B. (1969a). J. Molec. Biol. 40, 423-443.

Flamm, W.G. McCallum, M. and Walker, P.M.B. (1969b). J. Molec. Biol. 42, 441-455.

Ford, P.J. and Southern, E.M. (1973), Nature, New Biol. 241, 7-12.

Fry, K., Poon, R., Whitcombe, P., Idriss, J., Salser, W., Mazrimas, J. and Hatch, F. (1973). Proc. Natl. Acad. Sci. 70, 2642-2646.

Galibert, F., Sedat, J. and Ziff, E. (1974). J. Molec. Biol. 87, 377-407.

Gall, J. G. and Atherton, D.D. (1974). J. Molec. Biol. 85, 633-636.

Gall, J.G., Cohen, E.H. and Atherton, D.D. (1973). Cold Spring Harbor Symp. Quant. Biol. 38, 417-422.

Gall, J.G., Cohen, E.H. and Polan, M.L. (1971). Chromosoma, 33, 319-344.

Gall, J.G. and Pardue, M.L. (1969). Proc. Natl. Acad. Sci. 63, 378-383.

Gelderman, A.H., Rake, A.V. and Britten, R.J. (1971). Proc. Natl. Acad. Sci. 68, 172-176.

Georgiev, G.P. (1969). J. Theoret. Biol. 25, 473.

Gilbert, W. and Dressler, D. (1968). Cold Spring Harbor Symp. Quant. Biol. 33, 473-484.

Gillespie, D. and Spiegelman, S. (1965). J. Molec. Biol. 12, 829-842.

Graham, D.E. and Skinner, D.M. (1973). Chromosoma, 40, 135.

Guille, E. and Quetier, F. (1973). Prog. in Biophys. and Molec. Biol. 27, 123-142.

Hamkalo, B.A., Miller, O.L. and Bakken, A.H. (1973). Cold Spring Harbor Symp. Quant. Biol. 38, 915-919.

Harbers, K. and Spencer, J.H. (1974). Biochemistry, 13, 1094-1101.

Harbers, K., Harbers, B. and Spencer, J.H. (1974). Biochem. Biophys, Res. Commun. 58, 814-821.

Harbers, K., Harbers, B. and Spencer, J.H. (1975). Biochem. Biophys. Res. Commun. 66, 738-746.

Harel, J., Hanania, N., Tapiero, H. and Harel, L. (1968). Biochem. Biophys. Res. Commun. 33, 696-701.

Hatch, F.T. and Mazrimas, J.A. (1970). Biochim. Biophys. Acta, 224, 291-294.

Henderson, A. and Ritossa, F.M. (1970). Genetics, 66, 463-473.

Hennig, W., Hennig, I. and Stein, H. (1970). Chromosoma, 32, 31-63.

Hennig, W. and Walker, P.M.B. (1970). Nature, 225, 915-919.

Hood, L., Campbell, J.H. and Elgin, S.C.R. (1975). Ann. Rev. Genet. 9, 305-353.

Hörz, W., Hess, I. and Zachau, H.G. (1974). E.J.Biochem. 45, 501-512.

Hourcade, D. Dressler, D. and Wolfson, J. (1973a). Cold Spring Harbor Symp. Quant. Biol. 38, 537-550.

Hourcade, D., Dressler, D. and Wolfson, J. (1973b). Proc. Natl. Acad. Sci. 70, 2926-2930.

Hutton, J.R. and Wetmur, J.G. (1973a). Biochemistry, 12, 558-563.

Hutton, J.R. and Wetmur, J.G. (1973b). Biochem. Biophys. Res. Commun. 52, 1148-1155.

Ingle, J, Pearson, G.G. and Sinclair, J. (1973). Nature, New Biol. 242, 193-197.

Jensen, R.H. and Davidson, N. (1966). Biopolymers, 4, 17.

John, H., Birnstiel, M. and Jones, K. (1969). Nature 223, 582-587.

Jones, K.W. (1970). Nature 225, 912-915.

Jones, K.W. (1973). J. Medec. Genet. 10, 273-281.

Jones, K.W. and Corneo, G. (1972). Nature New Biol. 233, 268-271.

Jones, K.W., Prosser, J., Corneo, G., Ginelli, E. and Bobrow, M. (1973). In 'Constitutive Heterochromatin in Man; Modern Aspects of Cytogenetics', Symposia Medica Hoechst no. 6, pp.45-63. E.K.Schattaeur Verlag, Stuttgart.

Jones, K.W., Purdom, I.F., Prosser, J. and Corneo, G. (1974). Chromosoma, 49, 161-171.

Kedes, L. and Birnstiel, M. (1971). Nature, 230, 165-169.

Kelly, T.J. and Smith, H.O. (1970). J. Molec. Biol. 51, 393-409.

Kersten, W.H., Kersten, W. and Szybalski, W. (1966). Biochemistry, 5, 236-244.

Kimura, M. (1973). Cold Spring Harbor Symp. Quant. Biol. 38, 515-524.

Kimura, M. and Ohta, T. (1971), Nature, 229, 467-469.

Kit, S. (1961). J. Molec. Biol. 3, 711-716.

Kornberg, A., Bertsch, L.L., Jackson, J.F. and Khorana, H.G. (1964). Proc. Natl. Acad. Sci. 51, 315-323.

Kram, R., Botchan., M. and Hearst, J.E. (1972). J. Molec. Biol. 64, 103-117.

Kurnit, D.M., Shafit, B.R. and Maio, J.J. (1973). J. Molec. Biol. 81, 273-284.

Laird, C., McConaughy, B.L. and McCarthy, B.J. (1969). Nature, 224, 149-154.

Lee, C.H. and Wetmur, J.G. (1973). Biochem. Biophys. Res. Commun. 50, 879-885.

Lewin, B. (1974). Cell, 1, 107.

Lima-de-Faria, A. (1969). In 'Handbook of Molecular Cytology', (Lima-de-Faria, A. Ed.) pp.277-325. North-Holland, London.

Lima-de-Faria, A., Birnstiel, M. and Jaworska, H. (1969). Genetics Suppl. 61 (1), 145-159.

Lindsley, D.L. and Novitski, E. (1958). Genetics, 43, 790-798.

Ling, V. (1972). J. Molec. Biol. 64, 87-102.

Locker, D. (1976). Molec. Gen. Genet. 143, 261-268.

Loening, U.E. (1967). Biochem. J. 102, 251-257.

Macgregor, H.C., Horner, H., Owen, C.A. and Parker, I. (1973), Chromosoma, 43, 329-348.

Maniatis, T., Ptashne, M., Barrell, B.G. and Donelson, J.E. (1974). Nature, 250, 394-397.

Manteuil, S., Hamer, D.H. and Thomas, C.A. Jnr. (1975). Cell, 5, 413-422.

Marmur, J. and Doty, P. (1961). J. Molec. Biol. 3, 585.

Marmur, J., Round, R. and Schildkraut, C.L. (1963). Prog. in Nucleic Acid Res. 1, 231-300.

Mazrimas, J. and Hatch, F. (1970). Exp. Cell Res. 63, 462-466.

Mazrimas, J. and Hatch, F. (1972). Nature New Biol. 240, 102-105.

Marx, K.A. and Hearst, J.E. (1975). J. Molec. Biol. 98, 355-368.

McCarthy, B.J. and Farquar, M.N. (1972). In 'Evolution of Genetic

Systems', Brookhaven Symposium, (Smith, H.H. Ed.), pp.1-43, Gordon and Breach, New York, London and Paris.

McConaughy, B.L. and McCarthy, B.J. (1970). Biochem. Genetics, 4, 425-446.

Meselson. M. Stahl F.W. and Vinograd J. (1957), Proc. Natl. Acad. Sci. 43 581.

Mowbray, S.L., Gerbi, S.A. and Landy, A. (1975). Nature 225, 367-370.

Mowbray, S.L. and Landy, A. (1974). Proc. Natl. Acad. Sci. 71, 1920-1924.

Murray, K. (1973). Biochem. J. 131, 569-583.

Murray, K. (1974). In 'M.T.P. International Review of Science; Biochemistry, Series 1, vol.6', (Burton, K. Vol. Ed.) pp.1-60. Butterworths, London.

Murray, K. and Old, R.W. (1974), Prog. in Nucleic Acid Res. and Molec. Biol. 14, 117-185.

Nathans, D. and Smith, H.O. (1975). Ann. Rev. Biochem. 44, 273-293.

Ohno, S. (1972). In 'Evolution of Genetic Systems', Brookhaven Symposium, (Smith, H.H. Ed.), pp.366-370, Gordon and Breach, New York, London and Paris.

Ohta, T. and Kimura, M. (1971). Nature, 233, 118-119.

Paprooka, W. and Müller, W. (1974). Proc. Natl. Acad. Sci. 71, 699-703.

Pardue, M.L., Brown, D.D. and Birnstiel, M. (1973). Chromosoma, 42, 191-203.

Pardue, M.L. and Gall, J.G. (1970). Science, 168, 1356-1358.

Pardue, M.L. and Gall, J.G. (1975). In 'Methods in Cell Biology' (Prescott, D.M. Ed.), pp.1-16. Academic Press, New York and London.

Peacock, W.J., Brutlag, D., Goldring, E. Appels, R. Hinton, C.W. and Lindsley, D.L. (1973). Cold Spring Harbor Symp. Quant. Biol. 38, 405-416.

Peacock, A.C. and Dingman, C.W. (1967). Biochemistry, 6, 1818-1827.

Petersen, G.B. and Reeves, J.M. (1966). Biochim. Biophys. Acta, 129, 438-440.

Philippsen, P., Streeck, R.E. and Zachau, H.G. (1974). E. J. Biochem. 45, 479-488.

Philippsen, P., Streeck, R.E., Zachau, H.G. and Müller, W.G. (1975), E. J. Biochem. 57, 55-68

Procunier, J.D. and Tartof, K.D. (1975). Genetics. 81, 515-523.

Prescott, D.M., Bostock, C.J., Hatch, F.T. and Mazrimas, J.A. (1973). Chromosoma, 42, 205-213.

Prosser, J., Moar, M., Bobrow, M. and Jones, K.W. (1973). Biochim. Biophys. Acta, 319, 122-134.

Rae, P.M.M. (1970). Proc. Natl. Acad. Sci. 67, 1018-1025.

Rae, P.M.M. (1972). Adv. in Cell Molec. Biol. 2, 109-150.

Rice, N. (1971). Carnegie Inst. Sci. Yearbook, 69, 472-479.

Rice, N. (1972). In 'Evolution of Genetic Systems', Brookhaven Symposium, (Smith, H.H. Ed.), pp.44-79, Gordon and Breach, New York and London.

Rice, N. and Straus, N.A. (1973). Proc. Natl. Acad. Sci. 70, 3546-3549.

Ritossa, F.M. (1972), Nature New Biol. 240, 109-111.

Ritossa, F.M., Atwood, K.C., Lindsley, D.L. and Spiegelman, S. (1966a). In 'National Cancer Inst. Monograph, 23 pp.449-499.' (Vincent, W.S. and Miller, O.L. Eds.).

Ritossa, F.M., Atwood K.C. and Spiegelman, S. (1966b), Genetics 54, 819-834.

Robertson, H.D., Barrell, B.G., Weith, H.L. and Donelson, J.E. (1973), Nature New Biology. 241, 38-40.

Rosbash, M., Campo, M.S. and Gummerson, K.G. (1975). Nature 258, 682-686.

Sadowski, P.D. and Hurwitz, J. (1969). J. Biol. Chem. 244, 6192-6198.

Salomon, R., Kaye, A.M. and Herzberg, M. (1969). J. Molec. Biol. 43, 581-592.

Salser, W. (1974). Ann. Rev. Biochem. 43, 923-965.

Salser, W., Fry, K., Brunk, C. and Poon, R. (1972). Proc. Natl. Acad. Sci. 69, 238-242.

van der Sande, J.H., Loewen, P.C. and Khorana, H.G. (1972). J. Biol. Chem. 247, 6140-6148.

Sanger, F. (1975). Proc. Roy. Soc. (London), B. 191, 317-333.

Sanger, F., Brownlee, G.G. and Barrell, B.G. (1965). J. Molec. Biol. 13, 373-398.

Sanger, F. and Brownlee, G.G. (1967). In 'Methods in Enzymology', (Grossman, L. and Moldave, K. Eds.) vol.XIIa, pp.361-381, Academic Press, New York and London.

Sanger, F. and Coulson, A.R. (1975). J. Molec. Biol. 94, 441-448.

Sanger, F., Donelson, J.E., Coulson, A.R., Kössel, H. and Fischer, D. (1973). Proc. Natl. Acad. Sci. 70, 1209-1213.

Schildkraut, C.L. and Maio, J.J. (1968). Biochim. Biophys. Acta, 161, 76-93.

Schmitz, K.S. and Schurr, J.M. (1972). J. Phys. Chem. 76, 534-545.

Sederoff, R., Lowenstein, L. and Birnboim, H.C. (1975). Cell, 5 183-194.

Skinner, D.M. (1967). Proc. Natl. Acad. Sci. 58, 103-106.

Skinner, D.M. and Beattie, W.G. (1973). Proc. Natl. Acad. Sci. 70, 3108-3112.

Skinner, D.M., Beattie, W.G., Blattner, F.R., Stark, B.P. and Dahlberg, J.E. (1974). Biochemistry, 13, 3930-3937.

Skinner, D.M., Beattie, W.G., Kerr, M.S. and Graham, D.E. (1970). Nature, 227, 837.

Smith, G.P. (1973). Cold Spring Harbor Symp. Quant. Biol. 38, 507-514.

Smith, G.P. (1976). Science, 191, 528-535.

Smith, H.O. and Wilcox, K.W. (1970), J. Molec. Biol. 51, 379-391.

Southern, E.M. (1970), Nature, 227, 794-798.

Southern, E.M. (1971), Nature New Biol. 232, 82-83.

Southern, E.M. (1972), In 'Constitutive Heterochromatin in Man; Modern Aspects of Cytogenetics', Symposia Medica Hoechst no.6, pp.19-29, E.K. Schattaeur Verlag, Stuttgart.

Southern, E.M. (1974a). In 'M.T.P. International Review of Science, Biochemistry, Series One, vol. 6,' (Burton, K. Vol. Ed.), pp.101-139, Butterworths, London.

Southern, E.M. (1974b). Analyt. Biochem. 62, 317-318.

Southern, E.M. (1975). J. Molec. Biol. 94, 51-69.

Southern, E.M. and Mitchell, A.R. (1971). Biochem. J. 123, 613-617.

Southern, E.M. and Roizes, G. (1973). Cold Spring Harbor Symp. Quant. Biol. 38, 429-434.

Sparrow, A.H., Price, H.J. and Underbrink, A.G. (1972). In 'Evolution of Genetic Systems', Brookhaven Symposium, (Smith, H.H. Ed.), pp.451-494, Gordon and Breach, New York, London and Paris.

Sueoka, N. (1961). J. Molec. Biol. 3, 31-40.

Sutton, W.D. (1972). Nature New Biol. 237, 70-71.

Sutton, W.D. and McCallum, M. (1971). Nature New Biol. 232, 83-85.

Sutton, W.D. and McCallum, M. (1972). J. Molec. Biol. 71, 633-656.

Székely, M. and Sanger, F. (1969). J. Molec. Biol. 12, 394-409.

Tartof, K.D. (1973). Cold Spring Harbor Symp. Quant. Biol. 38, 491-500.

Tartof, K.D. (1975). Ann. Rev. Genet. 9, 355-386.

Tobia, A., Schildkraut, C.L. and Maio, J.J. (1970). J. Molec. Biol. 54, 499-515.

Vinograd, J., Morris, J., Davidson, N. and Dove, W.F. (1963). Proc. Natl. Acad. Sci. 49, 12-17.

Walker, P.M.B. (1968). Nature, 219, 228

Walker, P.M.B. (1969). Prog. in Nucleic Acid Res. and Molec. Biol. 9, 301

Walker, P.M.B. (1971a). Prog. in Biophys. and Mol. Biol. 23, 145-190.

Walker, P.M.B. (1971b). Nature, 229, 306-338.

Walker, P.M.B., Flamm, W.G. and McClaren, A. (1969). In 'Handbook of Molecular Cytology', (Lima-de-Faria, A. Ed.), pp.52-66. North-Holland, Amsterdam, London.

Walker, P.M.B. and McClaren, A. (1965). Nature, 208, 1175-1179.

Wallace, M.E. (1958). Phil. Trans. Royal Soc. London, B. 241, 211.

Waring, M. and Britten, R.J. (1966). Science, 154, 791-794.

Wellauer, P.K. and Dawid, I.B. (1974). J. Molec. Biol. 89, 379-395.

Wellauer, P.K., Reeder, C.H., Carroll, D., Brown, D.D., Deutch, A. Higashinakagawa, T. and Dawid, I.B. (1974). Proc. Natl. Acad. Sci. 71, 2823-2827.

Wells, R.D. and Ingle, J. (1970). Plant Physiol. 46, 178-179.

Wells, R.D., Learson, J.E., Grant, R.C., Shortle, B.E. and Cantor, C.R. (1970). J. Molec. Biol. 54, 465-497.

Wetmur, J.G. and Davidson, N. (1968). J. Molec. Biol. 31, 349-370.

Whitcombe, P., Fry, K. and Salser, W. (1974). In 'Methods in Enzymology', (Grossman, L. and Moldave, K. Eds.), vol. XXIX, pp.295-321, Academic Press, New York and London.

Wimber, D.E. and Steffensen, D.M. (1970). Science 170, 639-641.

Yasmineh, W.G. and Yunis, J.J. (1969). Biochem. Biophys. Res. Commun. 35, 779-782.

Yasmineh, W.G. and Yunis, J.J. (1970). Exp. Cell. Res. 59, 69-75.

Yoshimori, R.N. (1971). Ph.D. Thesis, University of San Francisco.

Yunis, J.J. and Yasmineh, W.G. (1970). Science, 168, 263-265.

Yunis, J.J. and Yasmineh, W.G. (1971). Exp. Cell. Res. 64, 41-48.

Yunis, J.J. and Yasmineh, W.G. (1972). Adv. in Cell. and Mol. Biol. 2, 1-46.

Ziff, E.B., Sedat, J.W. and Galibert, F. (1973). Nature New Biol. 241, 34-37.

Zimmer, C., Reinert, K.E., Luck, G., Wähnert, U., Löber, G. and Thrum, H. (1971). J. Molec. Biol. 58, 329-348.